THE UNIVERSITY *of* EDINBURGH

Cross-Lingual Genre Classification

Philipp Petrenz

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2014

Abstract

Automated classification of texts into genres can benefit NLP applications, in that the structure, location and even interpretation of information within a text are dictated by its genre. Cross-lingual methods promise such benefits to languages which lack genre-annotated training data. While there has been work on genre classification for over two decades, none has considered cross-lingual methods before the start of this project. My research aims to fill this gap. It follows previous approaches to mono-lingual genre classification that exploit simple, low-level text features, many of which can be extracted in different languages and have similar functions. This contrasts with work on cross-lingual topic or sentiment classification of texts that typically use word frequencies as features. These have been shown to have limited use when it comes to genres. Many such methods also assume cross-lingual resources, such as machine translation, which limits the range of their application. A selection of these approaches are used as baselines in my experiments.

I report the results of two semi-supervised methods for exploiting genre-labelled source language texts and unlabelled target language texts. The first is a relatively simple algorithm that bridges the language gap by exploiting cross-lingual features and then iteratively re-trains a classification model on previously predicted target texts. My results show that this approach works well where only few cross-lingual resources are available and texts are to be classified into broad genre categories. It is also shown that further improvements can be achieved through multi-lingual training or cross-lingual feature selection if genre-annotated texts are available in several source languages. The second is a variant of the label propagation algorithm. This graph-based classifier learns genre-specific feature set weights from both source and target language texts and uses them to adjust the propagation channels for each text. This allows further feature sets to be added as additional resources, such as Part of Speech taggers, become available. While the method performs well even with basic text features, it is shown to benefit from additional feature sets. Results also indicate that it handles fine-grained genre classes better than the iterative re-labelling method.

Acknowledgements

First and foremost, I would like to thank my two supervisors, Bonnie Webber and Victor Lavrenko, for their outstanding support throughout the duration of this PhD project. They have given me valuable feedback, advice, and guidance on countless occasions and I strongly benefited from their expertise and experience. Without their help, this project would not have been possible.

I also am very grateful to my PhD examiners, Iain Murray and Hinrich Schütze, for the time and effort they put into the examination process and for their excellent feedback during the viva.

I would further like to thank several members of the School of Informatics, who provided helpful comments in early stages of this project. Mirella Lapata, Maria Wolters, Benjamin Rosman, Annie Louis, and Ben Allison, in particular, have generously contributed advice and suggestions. I would also like to express my appreciation for the feedback from several anonymous reviewers who took the time to peer-review my submissions to journals and conferences.

I thank my family and friends for all the support and encouragement. Finally, I thank my girlfriend Helle for being fantastic.

This PhD project was partly financed by a Google Research Award, which made the work reported here possible.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Philipp Petrenz*)

Table of Contents

Chapter 1

Introduction

Using machine learning to automatically classify texts into categories has become a standard practice with a wide range of applications in Natural Language Processing (NLP), Information Retrieval (IR), and other fields. The nature of the data sets it apart from many other classification tasks. The domain of text involves specific characteristics and challenges, including its heterogeneity and the potential size of the feature space. Typically, even small corpora consist of tens or hundreds of thousands of unique words (Yang and Pedersen 1997). Traditionally, researchers such as Joachims (1998) and Sebastiani (2002) have focused on the topic of a text for this task and topical classification remains the most common to this day. However, a text is characterized not only by its topic. Other classification criteria have evolved, including sentiment (e.g., Pang et al. 2002), authorship (e.g., De Vel et al. 2001; Stamatatos et al. 2000b), and author personality (Oberlander and Nowson 2006), as well as categories relevant to filter algorithms (e.g., spam or appropriateness for different age groups).

Genre is yet another characteristic of a text, often described as orthogonal to topic. It has been shown by Biber (1991), and others after him, that the genre of a text has an impact on its formal properties. Therefore, it is possible to extract cues (e.g. lexical, syntactic, structural cues) from a text and use them as features to predict its genre. While work on genre classification has been reported for almost two decades, research efforts have increased significantly in the past ten years. Like in most other fields related to computational linguistics, academics thus far have focused on English texts. Recently however, genre classification has been carried out on a variety of other languages, as annotated non-English texts have become available and language-independent methods have become more popular.

This dissertation documents my PhD research on methods for genre classification

that can be used across languages. Broadly speaking, the idea is to train a classification model using genre annotated documents in one or more languages and subsequently use this model to predict text genres in another language, for which no labelled data is assumed to exist. This chapter aims to explain the term *genre* and to summarize prior work on genre classification, focusing on language-independent methods. As there is no substantial prior work on cross-lingual genre classification (CLGC), approaches from other fields like topic classification and sentiment analysis are reviewed, including a discussion of how suitable such methods are for the task of genre classification. This puts the project into perspective and motivates the methods and experiments described in the following chapters.

1.1 What are text genres and why do they matter?

At first glance, the meaning of the term *genre* seems obvious and intuitive. However, unlike the other text characteristics mentioned before (topic, authorship, etc.), the concept of genre is not clearly defined and a wide range of descriptions can be found in the literature. Many authors have introduced related and overlapping terms like *register*, *text type*, or *style* and the usage of such terms is far from consistent. Lee (2001) has compiled an excellent overview of the most common definitions, of which a subset is presented here.

In 1991, Biber defined genre categories solely by external criteria, that is a group of texts with a common communicative purpose (instructive, narrative, etc.). While he also showed that genres have an impact on linguistic features (e.g., the use of past tense), he argued that texts of the same genre were not necessarily coherent in their linguistic characterizations. Biber defined groupings based on linguistic form, or internal criteria, as text types. Swales (1990) also focused on the communicative purpose as the crucial criterion, but argued that texts of the same genre exhibit similar structures, styles, contents, and intended audiences as well. Lee (2001) himself advocates a more dynamic view of genres, which he sees as socially and culturally constituted text categories. He argues that genres have changed in terms of the *registers* they are associated with, which he defines as static instantiations of a conventionalised, functional configuration of language. In order to carry out genre classification experiments, many academics have adopted a more simplified and practical definition. For example, Kessler et al. (1997) see genres as extensible collections of text with shared communicative purposes (external) and common formal cues (internal).

Another unclear point is the distinction between genres and topics. Many text corpora are structured in categories blurring the boundaries between function and content. While Biber (1991) included topics in the definition of genres, many authors (e.g., Meyer zu Eissen and Stein 2004; Finn and Kushmerick 2006) have taken the stance that genre is orthogonal to topic. Others, like Karlgren and Cutting (1994) note that co-variance exists and some, like Kim and Ross (2008), are doubtful as to how well the two concepts can be separated in practice. Using a large 20-year corpus of news data, Petrenz and Webber (2011) have shown that genre and topic do in fact correlate strongly and that this correlation can change over time. However, it was also shown that by using appropriate features, it is possible to identify the genre of a text almost completely independently of its topic (Petrenz 2009).

It is beyond the scope of this project to discuss the definition of genre in great depth. However, as the methods proposed here use machine learning techniques to predict genre categories from extractable text features, a definition which includes internal criteria is appropriate. At the same time, external criteria are equally important, as there is no benefit in classifying texts based on linguistic features, unless the resulting categories share a communicative purpose. Therefore, even though it may be fairly broad, the definition of Kessler et al. (1997) above is adopted here.

Furthermore, genres are not regarded as atomic entities, but rather as a combination of different facets, as Kessler et al. refer to them: "A facet is simply a property which distinguishes a class of texts that answers to certain practical interests, and which is moreover associated with a characteristic set of computable structural or linguistic properties [...]" (Kessler et al. 1997). This is similar to the multi-dimensional framework of Biber (1995, see Section 1.2), in that genres are not defined along a single dimension.

This is important, as some genres may be identified by a certain facet, while for others, the same facet is irrelevant. For example, whether the author is trying to persuade the reader may only play a subordinate role in determining whether a text is a letter. It is however a crucial facet of an advertisement. No fixed definition is attempted here about which facets should be considered part of the definition of genre. However, as each facet is associated with certain conventions, which in turn dictate linguistic and structural choices, the same effect should be observable for internal criteria. That is, some genres should correlate with certain types of features, while others do not, or less so. While this is not the focus of this project, the experimental findings in Chapter 7 provide some evidence for this definition.

Beyond that, a very practical stance was adopted for this project. Genre-annotated

data in multiple languages is difficult to find, which dictated a fairly opportunistic sampling of genre palettes. However, the experiments and analyses use genre classes which are widely accepted and have been used in similar work before.

Further, more detailed discussion about the definition of genres, albeit with a focus on web genres, can be found in (Mehler et al. 2010), in particular in Chapters 1-4 (Santini et al. 2011; Karlgren 2011; Rosso and Haas 2011; Crowston et al. 2011) and Chapter 15 (Bruce 2011).

Genre classification can directly benefit information retrieval applications (Karlgren and Cutting 1994; Kessler et al. 1997; Finn and Kushmerick 2006; Freund et al. 2006), where users may want documents that serve a particular communicative purpose, in addition to their topic. For example, a web search for a topical keyword like *crocodiles* may return an encyclopaedia entry, a biological fact sheet, a news report about attacks in Australia, a blog post about a safari experience, a fiction novel set in South Africa, or a poem about wildlife. Most of these genres will be irrelevant to the user. Having classified indexed texts by genre would allow a search engine to provide additional selection criteria to reflect this. This was demonstrated to be beneficial empirically by Vidulin et al. (2007), who showed that restricting results to a genre category based on automatic classification yields better results than using keywords alone, even when these keywords are adapted to the target genre. A similar approach was used for a practical application by Stein et al. (2011), who developed a Firefox add-on called *WEGA* (Web-based Genre Analysis). This software annotates the snippets of search engine result pages with genre labels based on an automated classification of the linked websites. Rather than employing a strict filter, this provides the user with additional information about a websites, thus potentially making it easier to chose or reject a result.

Furthermore, genre classification can benefit language technology indirectly, where differences in the functional cues that correlate with genre may impact system performance. For example, Petrenz and Webber (2011) found that, within the New York Times corpus (Sandhaus 2008), the likelihood of the word *states* being a verb is considerably higher in letters (approx. 20%) than in editorials (approx. 2%). Part-of-Speech (PoS) taggers or statistical machine translation systems trained on a corpus of editorials and used for tagging or translating a corpus of letters to the editor might benefit from such knowledge. Kessler et al. (1997) mention that parsing and word-sense disambiguation can also benefit from genre classification and Webber (2009) found that genres have an impact on the distribution of discourse relations. Similarly, Pivovarova et al. (2013) show that the structure of events in a text depends on its genre. Knowledge about

this structure can be exploited by Information Extraction systems, as Pivovarova et al. (2013) propose for future work. Stubbe (2006) suggests that knowing the genre of a text could also improve automated summarization algorithms, as genre conventions dictate the location and structure of important information within a document. This is supported by more recent work in the field. Goldstein et al. (2007), for example, use a genre classifier to inform a text summarization system and conclude that "summaries that reflect a user's information seeking needs requires genre oriented goal-focused summarization." Their results show that a genre-specific summary outperforms a generic summary for most of the genres tested. Similarly, the work of Yatsko et al. (2010) exploits the differences in artistic, newspaper, and scientific texts to build genre-specific summarization algorithms.

1.2 Recent developments in genre classification

As in any other machine learning task, experiments on genre classification are mainly characterized by four criteria: The data, the input variables, the target variables, and the learning algorithm used. Of these, the last is probably the one that is discussed the least among researchers. Although various methods have been used and sometimes compared, most authors in the field of genre classification concentrate on text features rather than learning algorithms, at least for the time being. Among those used are discriminant analysis (e.g., Karlgren and Cutting 1994), C4.5 decision trees (e.g., Finn and Kushmerick 2006), Naïve Bayes (e.g., Lee and Myaeng 2002), k-nearest-neighbours (e.g., Wolters and Kirsten 1999) and neural networks (e.g., Kessler et al. 1997). However, the most popular method in recent years are support vector machines (Freund et al. 2006; Sharoff 2007; Kim and Ross 2008; Wu et al. 2010), which have been shown to work well on text classification tasks in general (Joachims 1998).

While experimenting with different learning algorithms is relatively easy, finding suitable data for genre classification experiments is not. Partly stemming from the controversial definition of genre discussed above, many different genre-annotated corpora and text collections have been used by researchers in the past. These corpora differ dramatically in size, topicality and origin, as well as in the natures and numbers of the genre categories they contain. Sharoff et al. (2010) found that the corpora usually used for genre classification experiments are not comparable to each other and that none of them can be seen as representative on its own. This obviously makes it hard to compare different methods. There are efforts to compile unified corpora to evaluate

genre classification methods on (Meyer zu Eissen and Stein 2004; Rehm et al. 2008; Berninger et al. 2008), but the situation remains unsatisfying. An interesting alternative approach was suggested by Sharoff (2006; 2010), who compiled corpora in several languages by downloading texts from the internet and annotating them using a semi-supervised classification method. It remains to be seen whether this approach will be adopted by the research community, as the method raises questions on the validity of the annotated genre labels.

Directly connected to the problematic data situation is a very heterogeneous choice of target variables, that is genre classes to predict, throughout the literature. Results have been reported for as little as two (Karlgren and Cutting 1994) and as many as 70 genres (Sharoff et al. 2010). The majority of publications treat genre classification as a single label problem, where each document belongs to exactly one of a limited set of known genres. The facets approach by Kessler et al. (1997) is more flexible, in that it can predict formerly unseen genres. More recently, Santini et al. (2006) proposed a method which assigns every document to either zero, one, or multiple genres and Chaker and Habib (2007) suggested a combination of different classifiers, each of which assigns a document to all known genre classes with different weights.

However, the most-often discussed aspect in the field of genre classification is the choice of input variables, that is text features. This choice is what sets it apart from topical classification, which typically exploits the frequency of content words, that is some variant of the bag of words representation.

Early approaches to genre classification proposed by Karlgren and Cutting (1994), Kessler et al. (1997), and Argamon et al. (1998) all rely on (partly) hand-crafted feature sets, which are specific to texts in English. They include counts and ratios of function words such as *we* or *therefore*, selected PoS tag frequencies, punctuation cues and other statistics derived from intuition or text analysis. Similarly language specific feature sets were later explored for mono-lingual genre classification experiments using German (Wolters and Kirsten 1999) and Russian (Braslavski 2004) documents. While such features are frequently reported to yield high prediction accuracies and at least some tend to be stable in the face of topical changes (Petrenz 2009), they are often selected to fit the genre palette used in the respective publication and might have to be adapted for different document collections. The facets approach by Kessler et al. (1997) is the notable exception, in that it — in theory — allows formerly unseen genres to be detected, although no such experiments were reported by the authors.

In the following years, automatically generated feature sets have become more pop-

ular. Most of these tend to be naturally language-independent, that is they might work in mono-lingual genre classification tasks in languages other than English. Examples are the word frequency based approaches suggested by Stamatatos et al. (2000a) and Freund et al. (2006), the image features suggested by Bagdanov and Worring (2001) and Kim and Ross (2008), the PoS histogram frequency approach by Feldman et al. (2009), and the character n-gram approaches proposed by Kanaris and Stamatatos (2007) and Sharoff et al. (2010). All of them were tested exclusively on English texts. While language independence is a popular argument and often claimed by authors, few have empirically shown that this is true for their approach. One of the reasons is the lack of appropriate corpora in languages other than English.

One of the few authors to carry out genre classification experiments in more than one language is Sharoff (2007). He uses document collections gathered from Internet search engines. Employing PoS 3-grams and a variation of common word 3-grams as feature sets, Sharoff classifies English and Russian documents into genre categories. However, while the PoS 3-gram set yields a respectable prediction accuracy for English texts, in Russian documents, no improvement over the baseline of choosing the most frequent genre class was observed. Another genre classification study with experiments in more than one language was carried out by Lee and Myaeng (2002). They exploit genre and topic labels in their training data to identify words that vary strongly across genres, but little across topics. Lee and Myaeng empirically show that such words make good features in a subsequent genre classification task, both for English and for Korean texts. One downside to their method is that it relies on a sufficiently large set of texts, annotated with both genre and topic categories. The data for the experiments reported by Lee and Myaeng (2002) was downloaded from the internet and annotated manually.

A different experimental setup was employed by Scholl et al. (2009). To classify web documents as one of five different genres, the authors exploit HTML tag frequencies, URL properties, text/mark-up ratios and CSS (Cascading Style Sheets) rule counts in addition to lexical and structural text features. Like Sharoff, they compile a corpus themselves. However, their approach is poly-lingual: Both the training set and the test set contain documents written in several European and Asian languages. The distribution is heavily skewed however, with 65% of texts in English and "about twenty" unspecified languages represented by only 21% of the data. No indication is provided of the distribution of genre classes and languages. The reported classification results are impressive. Unfortunately, there is no information on how well genres were predicted for each of the languages. Therefore, it is hard to say how language-independent

the approach actually is. What is more, the classifier used by Scholl et al. heavily relies on the structure dictated by the mark-up language, as they show by means of the information gain ranking of their features. While such features are an obvious choice in web document genre classification, they are unavailable in texts from non-digital sources or web documents in formats other than HTML.

To date, very few experiments have convincingly shown a successful language-independent approach to genre classification. Taking this one level further, there has been no work that as much as attempted *cross-lingual* genre classification prior to the start of this project. This is somewhat surprising, as similarities and differences in genre variations across languages were discussed almost 20 years ago in an extensive study by Biber (1995). Comparing the communicative purposes and linguistic features of different genres in English, Somali, Korean, and Nukulaelae Tuvaluan (a Polynesian language), he reached the conclusion that individual features, such as counts of adverbs, "do not provide a reliable basic for cross-linguistic generalizations". This may be for several reasons, including the fact that the same linguistic feature can be associated with a different communicative purpose in two different languages. Instead, Biber proposes a multi-dimensional approach, where factor analysis is used on a set of linguistic features to identify dimensions (or factors), which he claims are a more reliable way of comparing genres across languages. As the study shows, some of the dimensions have a striking similarity across languages.

Manually mapping the factors to communicative functions, Biber was able to show that dimensions reflecting interactiveness, production circumstances, informational focus, personal stance, and narration can be derived for all four languages. Moreover, he was able to show that the equivalent genres are very often placed in similar regions along equivalent dimensions. For example, editorials were uniformly identified as non-interactive, non-narrative and with an integrated informational focus. Although there are differences as well (e.g., letters are characterized as narrative in Korean, but non-narrative in Somali), this work shows that equivalent genres in different languages are similar in several ways and that this can be deduced from extractable text features. Biber (1995) does not carry out cross-lingual genre classification, but he shows that it is possible and provides an excellent starting point. However, Biber's approach relies heavily on prior knowledge of a language. The features he uses for factor analysis are hand crafted and derived from careful manual text analyses. Mapping dimensions from different languages based on similar communicative purposes is not an automated process either. This makes it less interesting, in particular for applications with more

than one target language or for work with poorly-resourced languages.

Recently, a study on cross-lingual genre classification for closely related languages was published by Snyman et al. (2012). Their research was partly inspired by early work on this project, which is described in Chapter 5 and was reported in (Petrenz 2012). The experiments by Snyman et al. (2012) predict coarse genre labels in Dutch texts, using a classifier trained on an Afrikaans corpus. The authors use a simple word frequency feature set and a Naïve Bayes classifier. The first of two evaluated approaches simply tests the Afrikaans classification model on Dutch texts, that is the language gap is ignored. The intuition is that the vocabulary overlap for closely related languages may be enough to yield good results. The second approach uses machine translation to translate target language texts from Dutch to Afrikaans before predicting genre labels, using the same classifier as before. This is effectively the baseline method of (Petrenz 2012). Snyman et al. (2012) show that their approach of using a word frequency based classifier directly across languages outperforms a random guess classification. Predictably however, the classifier performs better when texts are translated. Furthermore, their method is restricted to languages with a significant vocabulary overlap. Unfortunately, the work reported by Snyman et al. (2012) also has weaknesses in the evaluation of the proposed methods. For example, precision values are compared to the results of Bel et al. (2003), who report a different metric (accuracy), on a different set of texts written in different languages (English and Spanish), with a different number of classes (twelve) for a different task (topical classification). Furthermore, some of the results reported in (Snyman et al. 2012) are inconsistent. For example, the confusion matrix shown does not explain the reported precision and recall values, which in turn do not match with the claimed F1-Score. In combination with a questionable and opaque experimental set-up, these factors regrettably make the contribution less helpful, even for cross-lingual genre classification tasks in closely related languages.

1.3 Cross-lingual approaches

Apart from the work of Snyman et al. (2012) described above, there is no research on cross-lingual genre classification beyond this project. However, cross-lingual methods have been proposed for other text classification tasks. The first to report such experiments were Bel et al. (2003), who predicted text topics in Spanish and English documents, using one language for training and the other for testing. Their approach was to train a classifier on language A, using a document representation containing only

content words (nouns, adjectives and verbs with a high corpus frequency). These words were then translated from language B to language A, so that documents in language B could be represented in the same way and the classifier could make predictions.

Thereafter, cross-lingual text classification has typically been regarded as a *domain adaptation* problem (for a description, see Blitzer et al. 2006) and researchers have tried to exploit large sets of unlabelled data and/or small sets of labelled data in the target language. For instance, Rigutini et al. (2005) proposed an EM based algorithm, where the labelled documents are translated from the source language to the target language, before a classifier is trained and used to predict labels on a large, unlabelled set in the target language. These instances are then used to iteratively re-train the classification model and the predictions are updated until convergence occurs. Using information gain scores at every iteration to only retain the most predictive words and thus reduce noise, Rigutini et al. (2005) achieve a considerable improvement over the baseline accuracy, which is a simple translation of the training instances and subsequent mono-lingual classification. They, too, classified texts by topics and used a collection of English and Italian newsgroup messages. Similarly, researchers have used semi-supervised bootstrapping methods like co-training (Wan 2009) and other domain adaptation methods like structural correspondence learning (Prettenhofer and Stein 2010) to carry out cross-lingual text classification.

All of the approaches described above make use of machine translation systems, even if some make an effort to keep translations to a minimum. This has several disadvantages however, as it makes their application dependent on the availability of parallel corpora, which may not always be available, in particular for poorly researched languages. It also introduces problems due to word ambiguity and morphology, especially where single words are translated out of context. A different method is proposed by Gliozzo and Strapparava (2006), who use latent semantic analysis on a combined collection of texts written in two languages. The rationale is that named entities such as *Microsoft* or *HIV* are identical in different languages. Exploiting the correlation of terms, the algorithm can identify semantically similar words in both languages. The authors use these mappings to carry out cross-lingual topic classification, and their results are promising. However, additionally using bilingual dictionaries yielded a considerable improvement, as Gliozzo and Strapparava (2006) also report.

While all of the methods above could technically be used in any text classification task, the idiosyncrasies of genres pose additional challenges. Techniques relying on the automated translation of predictive terms (e.g., Bel et al. 2003; Prettenhofer and

Stein 2010) are workable in the contexts of topics and sentiment, as these typically rely on content words such as nouns, adjectives and adverbs. For example, a word like *hospital* is indicative of a text from the medical domain, and *excellent* occurs primarily in positive reviews, as opposed to negative ones. Such terms are relatively easy to translate, even if not always without uncertainty. Genres, on the other hand, are often classified using function words (e.g., Karlgren and Cutting 1994; Stamatatos et al. 2000a) like *of*, *it*, or *in*. Reliably translating such words out of context can be difficult or even impossible. This is true in particular if there are differences in morphology, since what are function words in one language may be morphological affixes in another. Besides, even if it was possible, it is unclear whether a high frequency of a certain function word in language A indicates the same communicative purpose as does its counterpart in language B.

Similarly, using the bilingual low-dimension approach by Gliozzo and Strapparava (2006) may work for genre classification in theory. However, it relies on certain words to be identical in two different languages. This is the case for named entities, which also indicate topic, rather than genre. A text containing the words *Obama* and *McCain* will almost certainly be about the U.S. elections in 2008, or at least about U.S. politics. On the other hand, there is little indication of what its genre might be: It could be a news report, an editorial, a letter, an interview, a biography or a blog entry, just to name a few. Because topics and genres correlate, one would probably reject some genres like instruction manuals or fiction novels. However, uncertainty is still large and as Petrenz (2009) showed, relying on such correlations can be dangerous. This is particularly true in the cross-lingual case, as it is not clear whether genres and topics correlate in similar ways in a different language.

1.4 Overview

This dissertation aims to support the following thesis.

> **While automated cross-lingual genre classification (CLGC) can benefit from extensive linguistic resources and/or machine translation, it requires neither. Instead, I demonstrate that CLGC can be effectively achieved with (1) a set of genre-annotated source language texts, (2) a set of unlabelled target language texts available at training time, to develop and improve the classifier, and (3) a sensible selection or weighting of simple features in both languages. Moreover, some of the presented methods can incorporate linguistic resources and machine translation, if available, to further improve results.**

In other words, this PhD project aims to exploit extractable text features to build cross-lingual genre classifiers. As there is no prior research on CLGC, little guidance is available about which features and which machine learning techniques work well for such a task. Therefore, this project aims to explore and compare several approaches. A particular focus is on answering the question of whether genres can be classified across languages if only a restricted amount of linguistic resources is available. That is, can satisfactory performances be achieved without the use of machine translation? Are simple features, some of which were mentioned earlier in this chapter and which do not require supervised parsers and taggers, predictive across languages? How can cross-lingual correlation with genre be quantified and used for sensible feature selection? Which machine learning techniques lend themselves for the task of CLGC? While this project cannot provide a final answer to all of these questions, I aim at providing empirical evidence to advocate the methods proposed in the later chapters of this dissertation.

Furthermore, it can be seen as a motivation and starting point for other researchers. To this end, Chapter 2 describes publicly available text corpora that I have identified and/or restructured to be usable for CLGC. They can be used for cross-lingual genre experiments and to develop, evaluate, and compare future methods. Chapter 3 then describes four broad types of text features that have been evaluated for this task. This includes formerly proposed features for mono-lingual genre classification, as well as new additions. This is hoped to inspire further work into further sets of features for the task of CLGC.

The experiments and methods of this project all work on the same assumptions. A sufficiently large set of genre-annotated texts is required in the source language. (While in practice, this would typically be English, other languages have been used in the experiments as well.) In addition, a set of unlabelled target language texts is assumed to be available at training time. While this is not absolutely necessary for a genre prediction, the classifiers exploit this data to improve their performances in a semi-supervised fashion. All methods can be used with a minimum of linguistic resources and knowledge about the target language. However, some approaches can incorporate additional resource-based knowledge, such as features obtained through machine translation. Note that the above does not apply to the baselines, many of which require cross-lingual resources.

The remainder of this document is structured as follows. Chapter 2 describes the text corpora used for the analyses and experiments of this project. This includes pre-

processing steps and corpus statistics. The features used by the different machine learning approaches are discussed and visualized in Chapter 3. Chapter 4 provides an overview of baselines that were used to compare the performances of different classification methods.

In Chapter 5, an iterative SVM algorithm based on the work in (Petrenz 2012) is presented and evaluated for the problem of CLGC. This exploits simple cross-lingual features to bridge the language gap and uses iterative re-labelling to improve performance. It differs from similar algorithms in separating feature sets for cross-lingual and target language specific learning, as well as in selecting texts with high label confidence for re-labelling.

Chapter 6 extends this work by evaluating the impact of training on texts of more than one language, based on (Petrenz and Webber 2012b). This assumes that genre-annotated text is available for multiple languages, but not for the target language. A classifier can then be trained on a multi-lingual training set and features can be evaluated and selected based on their cross-lingual predictive power.

A variant of label propagation is discussed as an alternative classification method in Chapter 7, based on work in (Petrenz and Webber 2012a). A multi-layer graph is proposed, which can exploit different types of features. Both source and target language text propagate their known or predicted genre labels through different channels, depending on what genre they belong to, or are believed to belong to. This is shown to work well for fine-grained classification tasks and where resources allow many different types of features.

The conclusion in Chapter 8 discusses advantages and disadvantages of all these approaches and Chapter 9 gives pointers to future work.

Note that the journal squib (Petrenz and Webber 2011) contains both work that was carried out as part of the (Petrenz 2009) MSc thesis and more recent work which is part of this project. In order to avoid any confusion, I refer to (Petrenz and Webber 2011) only for work done subsequent to (Petrenz 2009) in this document. Conversely, any work that was already submitted for the MSc thesis is cited as (Petrenz 2009), and should be viewed as prior work, rather than novel contributions.

Chapter 2

Corpora

This chapter introduces the corpora used for the experiments and gives details about their sizes and structures. The languages and genres represented will be highlighted and compared. In addition, methods and justification for pre-processing the texts are presented. Each section in this chapter (except Section 2.5) discusses a set of corpora which contain comparable genres and are used together in the experimental frameworks of this project.

As mentioned in Section 1.2, suitable text collections for genre classification are rare, even for mono-lingual tasks in English. The challenge of finding corpora for cross-lingual experiments is even bigger. This is because, at the very least, texts from two languages are required, with a comparable set of manually annotated genres for both. More than two languages would of course be preferable, as would be a large quantity of texts in each language, or further annotation. However, there are very few corpora that even fulfil the basic requirements. Therefore, an opportunistic approach was taken in that experiments were based on the genres and languages for which data was available, even where the respective corpora were sub-optimal in their sizes, genre distributions, annotation, or sampling.

All data that is discussed below is publicly available, although some corpora are offered for a fee. This ensures that all analyses and results presented in this document can be reproduced by anyone interested in the subject. It also means that future approaches to CLGC can be evaluated and compared based on this data.

Informative Texts	Press Texts	Reportage	(44 texts)
		Editorials	(27 texts)
		Reviews	(17 texts)
	Miscellaneous	Religion	(17 texts)
		Skills, Trades & Hobbies	(36 texts)
		Popular Lore	(48 texts)
		Biographies & Essays	(78 texts)
	Non-Fiction	Reports & Official Documents	(30 texts)
		Scientific Writing	(80 texts)
Imaginative Texts	Fiction	General Fiction	(29 texts)
		Mystery & Detective Fiction	(24 texts)
		Science Fiction	(6 texts)
		Adventure & Western Fiction	(29 texts)
		Romantic Fiction	(29 texts)
		Humor	(9 texts)

Table 2.1: Genres in the Brown corpus. Categories are identical in the LCMC, except Western Fiction is replaced by Martial Arts Fiction.

2.1 Brown, LCMC, and SUC

Most of the experiments in this project are based on data from the Brown Corpus (BC), the Lancaster Corpus of Mandarin Chinese (LCMC), and the Stockholm-Umeå-Corpus (SUC) written in English, Chinese, and Swedish respectively. These corpora are used to compare the performances of the iterative re-labelling (Chapter 5) and the label propagation (Chapter 7) methods to those of different baselines (Chapter 4).

The Brown University Standard Corpus of Present-Day American English (Francis and Kucera 1979) contains 500 texts, manually classified into a hierarchy of genres. Table 2.1 shows this hierarchy and the number of texts for each genre. Each text is a sample of approximately 2,000 words. The exact word count varies slightly, as only complete sentences are included. The BC has often been used for mono-lingual genre classification (e.g., Karlgren and Cutting 1994; Kessler et al. 1997), in particular in early studies on the subject.

The LCMC (McEnery and Xiao 2004) was constructed in a very similar way to the BC. Like its English counterpart, it comprises 500 text samples, which fall into the

same hierarchy of genres. The exception is the *Adventure & Western Fiction* category of the BC, which is replaced by *Martial Arts Fiction* in the LCMC for cultural reasons. The numbers of texts for each genre are similar, but not identical (see Figure 2.1).

The SUC (Gustafson-Capková and Hartmann 2006) is the Swedish equivalent to the BC, in that it also contains 500 text samples of approximately 2,000 words and a similar genre hierarchy. While the sampling process was based on that of the BC, the authors made a few changes to the categorization. Most importantly, there is no separate *Religion* category in the SUC. While there are religious texts represented in the corpus, these were classified as either *Skills, Trades, and Hobbies*, *Popular Lore*, or *Scientific Writing*. The reasoning was that *Religion* is in fact a topical category, rather than a genre: "No other main category is defined on the basis of its subject matter" (Gustafson-Capková and Hartmann 2006). Such concerns have previously been raised by academics working on genre classification with the BC (Kessler et al. 1997). In this light, the removal of such a category seems logical. Therefore, I ignored the texts from the *Religion* categories in the BC and LCMC for all experiments. This means that only 483 English and 483 Chinese texts were used.

Another change in the SUC is a different sub-categorization of the *Fiction* genre. Instead of the six categories in the BC and LCMC (see Table 2.1), only four categories are used: *General Fiction*, *Mysteries and Science Fiction*, *Light Reading*, and *Humour*. The experiments in this project take this reduction further and always keep the *Fiction* category intact, i.e. it is treated as an atomic unit. The reasoning is that in all three corpora, this genre differs in its sub-categories. The same strategy was used in the BC experiments of Sharoff et al. (2010). It is also in line with the work of Lee (2001) with the British National Corpus. Despite his very fine-grained categorisation of 70 genres, Lee established the class *Fiction Prose*, where some texts have keywords such as *General* (e.g. text ID: FET), *Mystery* (e.g. FU2), *Adventure* (e.g. EFJ), *Romance* (e.g. H9L), and *Humorous Prose* (e.g. ASD). While these texts would have fallen into different fiction sub-categories in the BC/LCMC and SUC, for many applications, *Fiction* might be a sufficiently fine-grained entity, without the need to further break it down.

Like earlier work on mono-lingual genre classification (e.g., Karlgren and Cutting 1994), I use all three layers of genre granularity shown in Table 2.1 for experiments. Due to the adaptations described above, the most fine-grained classification task uses nine genres, rather than the 15 in the BC. The other tasks use two and four categories respectively. Figure 2.1 illustrates the number of texts in each genre category for each

Figure 2.1: Distribution of texts in the BC (English, 483 texts), LCMC (Chinese, 483 texts), and SUC (Swedish, 500 texts) for classification tasks with different levels of granularity: Two genres (top left), four genres (top right), and nine genres (bottom). 17 texts were omitted from the BC and LCMC each, to adapt to the classification scheme of the SUC.

of these three levels of granularity.

The texts in all corpora, including the Chinese LCMC, are tokenized on word, sentence, and paragraph level by default. I used this annotation to derive values for simple cross-lingual features (see Chapter 3). No pre-processing was applied to the texts beyond the automated PoS tagging described in Section 2.5. Due to the sampling strategy of these corpora, it is possible that the first and/or the last paragraph of a text sample are incomplete, which may affect paragraph-based feature values. No adjustments were made to correct for this, as it was assumed that it would affect texts of different genres similarly in each language.

Some of the baselines require machine translation. To this end, I employed the Google Translate Research API[1] to translate all texts into the two languages they were not written in (e.g. a Chinese text would be translated into both English and Swedish). The reason for choosing Google Translate was the vast amount of data it incorporates, the wide range of supported language pairs, and the state-of-the-art technology, all of which should result in meaningful baseline performances. Note that Google Translate is an online service and results may change due to updated algorithms or resources. All the full text translations for this project were carried out between 26/08/2013 and

[1] http://translate.google.com

28/08/2013.

2.2 New York Times and tageszeitung

Some experiments are carried out on data from the New York Times Annotated Corpus (NYTAC) and the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z). The genres in these newspaper corpora can all be considered sub-categories of the *Press text* of the BC, although the meta-data here allows for a finer distinction than that described in Section 2.1. The classification task is therefore more restricted and fine-grained. The two corpora also add English-German as a language pair.

The NYTAC (Sandhaus 2008) comprises over 1.8 million English articles (1.1 billion words) published between 1987 and 2007. As its name suggests, it is annotated with an extensive amount of meta-data, indicating authorship, publication date, and topic, among other things. While it is not directly annotated, some of the meta-data fields allow inference about the genre of a text. For this project, I used an updated version of the genre assignment process described by Petrenz (2009). For easy reference, it is summarized here:

> "While there is no explicit meta-data tag for the genre of an article, an array of fields was found to be particularly useful for the purpose of this project. An example is the tag Taxonomic Classifier, which places a document into a hierarchy of articles. This is a structured mixture of genres and topics. A document can be classified in several such hierarchies. Throughout the corpus, 99.5% of documents contain this field, with an average of 4.5 taxonomic classifiers assigned to each article.
>
> Examples include:
> `Top/Features/Travel/Guides/Destinations/Europe/Turkey`
> `Top/News/Business/Markets`
> `Top/News/Sports/Hockey/National Hockey League/Florida Panthers`
> `Top/Opinion/Opinion/Letters`
>
> Another valuable field is the Types of Material tag. It specifies the editorial category of the article, which in some cases corresponds to the definition of genre used for this project. In total, 41.5% of the documents in the corpus have a Type of Material tag assigned to them. The values are typically exclusive, even though a negligible amount of documents with more than one tag exists. There is no fixed set of values or hierarchy as there is for the taxonomic classifiers. Also, the Type of Material fields often contain errors, misspellings or very specific information about an article.

> Examples include:
> `Obituary`
> `Letter`
> `Letterletter`
> `Editorial photo of homeless person`
>
> [...]
>
> As there is no News tag in the Types of Material field, the Taxonomic Classifier field was used to identify [...] categories as follows:
>
> News
> Taxonomic classifier begins with `Top/News` excluding
> `Top/News/Obituaries`
> `Top/News/Correction`
> `Top/News/Editors Notes`
>
> Review
> Taxonomic classifier is one of the following
> `Top/Opinion/Opinion/Editorials`
> `Top/Opinion/Opinion/Op-Ed Top/Opinion/Opinion/Op-Ed/***`
> `Top/Features/***/Columns Top/Features/***/Columns/***`
> `Top/Features/***/Reviews Top/Features/***/Reviews/***`
> where *** can be anything, including several sub-hierarchies.
>
> Letter
> Taxonomic classifier is `Top/Opinion/Opinion/Letters`
>
> [...]
>
> It was decided to use the Types of Material field as an additional filter. Documents were only classified as news articles if they fulfilled both the criteria mentioned above and contained no Types of Material tag. For the Review class, only documents which were tagged Review, Editorial or Op-Ed were taken into consideration. No additional constraints were required for the Letter class."

The work in (Petrenz 2009) required the distinction of three genre classes only. However, for this project, a more fine-grained palette of genres was required to match the categories in the *tageszeitung* corpus (see below). Therefore, the process was changed and the *Type of Material* field was used to establish the categories, rather than just as a filter. Using a threshold of at least 1,000 texts per genre class to avoid the long tail of questionable categories with little data, 18 distinct values were identified for the *Type of Material* tag (the few texts with more than one tag were not added to any

News Reportage	521,622	Letters	138,003	Paid Death Notices	132,026
Statistics	111,982	Reviews	110,533	Editorials	53,518
Corrections	47,707	Obituaries	37,135	Summaries	31,005
Captions	30,571	Biographies	17,740	Schedules	16,117
Lists	14,084	Questions	6,862	Analyses	6,346
Texts	3,841	Interviews	3,537	Series	3,281
Chronologies	1,063				

Table 2.2: Identified genres in the New York Times Annotated Corpus and number of texts in each category.

category). However, this does not include a *News Reportage* genre, like in the Brown and tageszeitung corpora. Since the data comes from the New York Times, reportage is of course part of the corpus – among the 58.5% of articles with no *Type of Material* tag. This genre class was therefore defined as any text with no such tag and with at least one *Taxonomic Classifier* starting with `Top/News`. Furthermore, documents that had *Taxonomic Classifier* values starting with `Top/Classifieds` or `Top/Opinion` were excluded. The resulting 19 categories and the distribution of texts for each are shown in Table 2.2.

While not all of these categories would necessarily be considered genres (e.g., *Texts*), most of them are sensible categories, and several have been used in genre classification tasks before. Of these, only a subset was used, as explained below. However, the identified set of genres in Table 2.2 could be a valuable starting point for future work on both mono-lingual and cross-lingual genre classification. Beyond the work reported in (Petrenz 2009) and (Petrenz and Webber 2011), the NYTAC has not yet been exploited for research in these fields.

Texts in the NYTAC are stored in XML files and the corpus includes Java libraries to extract meta-data and different text parts. I used this tool to extract the headline, abstract, byline, dateline, and body of a text, which were subsequently concatenated, separated by paragraph boundary markers. Furthermore, the following approach was adopted from Petrenz (2009):

> "Looking at the results, it was found that in many cases the lead paragraph had been automatically added to the text content. This led to redundant sentences, as illustrated below (sample taken from document 0000702.xml).
> `LEAD: New York City won its three-year fare freeze in Albany`

> last week, though from downstate the ice looked a little mushy. New York City won its three-year fare freeze in Albany last week, though from downstate the ice looked a little mushy. The Legislature voted [...]
>
> Therefore, any initial paragraph starting with LEAD: was removed before further processing."

While paragraph boundaries are given in the texts, sentence boundaries are not. They were therefore annotated using the unsupervised *Punkt* algorithm (Kiss and Strunk 2006) implemented in the NLTK (Bird et al. 2009) framework.

The TüPP-D/Z (Müller 2004; Ule 2004) is made up of texts from the German newspaper *die tageszeitung*, published between 1986 and 1999. At about 600,000 articles, it is significantly smaller than the NYTAC and is not as thoroughly annotated with meta-data. The corpus is categorized into seven genres (identifiable through the <AR> tag), namely *TAZ-Bericht* (news reportage; 381,701 texts), *Agentur* (news wire copy; 141,102 texts), *Kommentar* (commentary/editorial; 35,537 texts), *LeserInnenbrief* (letter to the editor; 17,755 texts), *Interview* (13,542 texts), *Dokumentation* (documentary; 10,979 texts), and *Portrait* (biography; 6,209 texts). Of these, news wire copies and documentaries are no separate categories in the NYTAC, so they were ignored for this project. On the other hand, the *Kommentar* category actually contains two genres found in the NYTAC: editorials and reviews. As there is no way to distinguish these from the meta-data in the TüPP-D/Z, a subset of 734 *Kommentar* texts were manually categorized as part of this project. Of these, 218 were found to be reviews and 516 were found to be editorials.

The TüPP-D/Z corpus is fully tokenized on word, sentence, and paragraph level, and the respective boundaries were used for this project. All available parts of the texts were used, including headlines and other auxiliary information.

The full list of texts in each genre category for both the NYTAC and the TüPP-D/Z is available online[2]. For the NYTAC, this is presented as lists of relative paths and file names, as the corpus contains a single file for each article. For the TüPP-D/Z, lists of file names and start/end positions are provided, as the corpus contains one file per day of publication, with multiple articles in each file. Note that, for copyright reasons, no textual data is included – the corpora need to be obtained separately.

For the experiments in Chapter 7, a balanced set of the six overlapping genres is used for texts from both corpora. The smallest classes in the NYTAC and TüPP-D/Z are interviews and reviews respectively, with 3,537 and 218 texts. From all other genres,

[2] http://homepages.inf.ed.ac.uk/s0895822/dissertation

an equal amount of texts was sampled randomly. The English source language and the German target language therefore comprise 21,222 and 1,308 texts respectively in total.

2.3 Reuters, Acquis, and Europarl

The experiments on multi-lingual training (see Chapter 6) required genre annotated texts written in more than three languages. However, no publicly available multi-genre corpora were found with a comparable categorization in several languages. It was therefore decided to use three multi-lingual corpora, each of which includes texts from a single genre written in several languages: the Reuters volume 1+2 corpus (Rose et al. 2002), the Europarl corpus (Koehn 2005), and the JRC-ACQUIS corpus (Steinberger et al. 2006). All three corpora contain a large number of texts in Danish, English, French, German, Italian, Portuguese, Spanish, and Swedish. (Although all three also contain texts in Dutch, there are comparatively few Dutch texts in the Reuters corpus, so Dutch texts were not used in the experiments.) The source corpora were re-organized to obtain a comparable corpus that contains texts in eight languages and three genres: newswire texts, transcribed speech, and legal texts. Note that the corpus is *comparable* since it contains texts from a fixed set of genres, but not necessarily topics.

Since the source corpora are in different formats, some pre-processing was necessary. The XML markup was removed from the Reuters newswire texts, and only the contents of the tags `<headline>`, `<byline>`, `<dateline>`, and `<text>` were kept. Paragraph markers were kept in the text. The texts in the Europarl corpus were divided up by speaker, that is, each speech was considered a distinct document. The `<speaker>` tags were then removed, but paragraph markers were kept. Missing speeches were ignored: The only requirement was that each text contains at least one token. The JRC-ACQUIS corpus comprises several subgenres within the legal domain, including treaties, agreements and proposals. Therefore, only documents from CELEX[3] sector 3 (legislation) were used, as this is the largest group within the corpus. The text within the `<body>` tags was extracted, again keeping the paragraph structure intact.

All texts were segmented into sentences using the unsupervised *Punkt* algorithm (Kiss and Strunk 2006) implemented in the NLTK (Bird et al. 2009) framework. Since Europarl and JRC-ACQUIS are parallel corpora, it was ensured that no translation of the same text was used in any two sets in our experiments. For Europarl texts, only

[3]CELEX (Communitatis Europeae Lex) is a database for European Union law documents. All texts in the JRC-ACQUIS corpus are classified by CELEX sector and document type.

the speaker's language (i.e., the language in which the original parliamentary speech was given, which is indicated in the meta-data) was used for any given text. For JRC-ACQUIS, the choice was random, since the corresponding journal is published in all European languages simultaneously.

A balanced distribution of genres and languages was used in the experiments with these corpora, that is the same number of texts from each genre and each language were selected. Splitting the legislation texts of the JRC-ACQUIS yielded 1,942 documents in each of the eight languages. To keep the genre distribution in the corpus balanced, 1,942 documents were randomly (with the above restrictions) sampled from both the Reuters and the Europarl corpora. The resulting eight sets each contained 5,826 texts from a single language. A list with identifiers of the texts used for experiments can be found online[4], along with scripts to extract and clean texts from the source corpora.

2.4 CIIL and BNC

The cross-lingual classification method presented in Chapter 7 requires little to no knowledge of the target language, if the respective resources are unavailable. Therefore, the Central Institute of Indian Language (CIIL) corpus was used to carry out experiments with target languages for which fewer such resources are available. The corresponding source language in these experiments was English, with data taken from the British National Corpus (BNC).

The CIIL corpus (Central Institute of Indian Languages 2011) contains texts from 10 Indian languages: Assamese (1,109 texts), Bengali (1,270 texts), Hindi (1,233 texts), Kannada (483 texts), Malayalam (598 texts), Marathi (465 texts), Oriya (1,220 texts), Punjabi (896 texts), Tamil (761 texts), and Telugu (776 texts). The corpus is not parallel, that is the texts are not translations of each other. Each text is stored in a separate file, with a preceding line of meta-data, which can include genre and topic markers, as well as author name, title, medium, publication year, and other information. Unfortunately, this annotation is inconsistent and can differ both from language to language and from text to text within a language. To illustrate this, consider the following two examples, taken from an Assamese text and a Tamil text respectively:

<APH02><Nat&Phy-Sc><Physics---><1990><Book-><আইনষ্টাই><শিবনাথ--><0443>

[4] http://homepages.inf.ed.ac.uk/s0895822/dissertation

<sarapo>Natural Sciences><Natural, Physical and Professional Sciences><Medicine/Ayurveda/Homeopathy><1988><Book><சரபோஜி மன்னரின் மருத்துவ முறைகள்><ஏன். ஜ?னார்த்தனன்><35>

Both texts belong to the *Natural, Physical, and Professional Sciences* genre. However, different tag names and positions make genre identification non-trivial. Still, through correcting spelling mistakes and creating a set of classification rules, most texts can be assigned to a genre category based on their meta-data. The categories I used in this project are biography, commerce, essay, child fiction, adult fiction, natural science, and social science. These seven genres are further grouped into three and two broad categories, as shown in Table 2.3.

After the initial line of meta-data, the texts are raw, that is no further mark-up or tags are provided in the corpus. However, the texts are formatted in slightly different ways. For example, some texts include page numbers as part of the text, while others do not. For this reason, I carried out experiments on Tamil and Malayalam texts only. The formatting in these two languages is comparatively consistent, and similar between the languages. Furthermore, Malayalam in particular is a poorly resourced language and no machine translation is available from either *Google Translate* or *Bing Translator*.

Some basic cleaning and pre-processing was necessary. Where page numbers were present in the text files, these were removed. Furthermore, sentence boundaries are not annotated in the corpus and had to be identified, since some of the features exploited by the cross-lingual classification methods rely on them. After discussing the problem with a Tamil native speaker, a very simple rule based approach was chosen, where sentences were defined to end after a full stop, question mark, or exclamation mark, except where a full stop was preceded by a number (e.g. *21.*). Furthermore, the line breaks in the corpus texts were removed, since they appear to be artefacts from prior processing, and have no correlation with sentence or paragraph boundaries.

The BNC (Burnard 2000) contains 4,054 English texts, of which 3,144 are written and 910 spoken. The written texts are further categorized along several dimensions, such as domain (e.g. *World Affairs*, *Arts*), medium (e.g. *Book*, *Periodical*), and target audience (e.g. *Children*, *Teenager*). The corpus does not have a proper genre categorisation, however. To remedy this, Lee (2001) manually identified 70 genres in the BNC, of which 46 are for written text, and assigned each text in the corpus to one of these categories. As can be expected from the high number of genre classes, the categorization is very fine-grained. For example, personal letters are distinguished from professional letters, as are school essays from university essays.

		Genre	Rule for BNC texts
Fiction and Essays		Essay	All texts tagged W_essay_school or W_essay_univ. Also text IDs A05, A0T, ASK, CK1, FB4, FYX, GXK, HDB, KAN, and KAL (all are described as *Essay(s)* by Lee or are titled as such and belong to none of the other six genres).
		Children Fiction	All texts tagged W_fict_prose with audience age child.
		Adult Fiction	All texts tagged W_fict_prose with audience age adult.
Non-Fiction	Other	Biography	All texts tagged W_biography, except text ID KAM (also tagged Essay in Lee's description).
		Commerce	All texts tagged W_commerce.
	Scientific Writing	Natural Science	All texts tagged W_ac_nat_science, W_ac_tech_engin or W_ac_medicine. Three texts with alternative genre tags (e.g. Also W_ac_soc_science) were removed.
		Social Science	All texts tagged W_ac_soc_science or W_ac_polit_law_edu. 46 texts with alternative genre tags (e.g. Also W_commerce) were removed.

Table 2.3: Rules for assigning BNC texts to one of seven genre classes, based on the labelling by (Lee 2001).

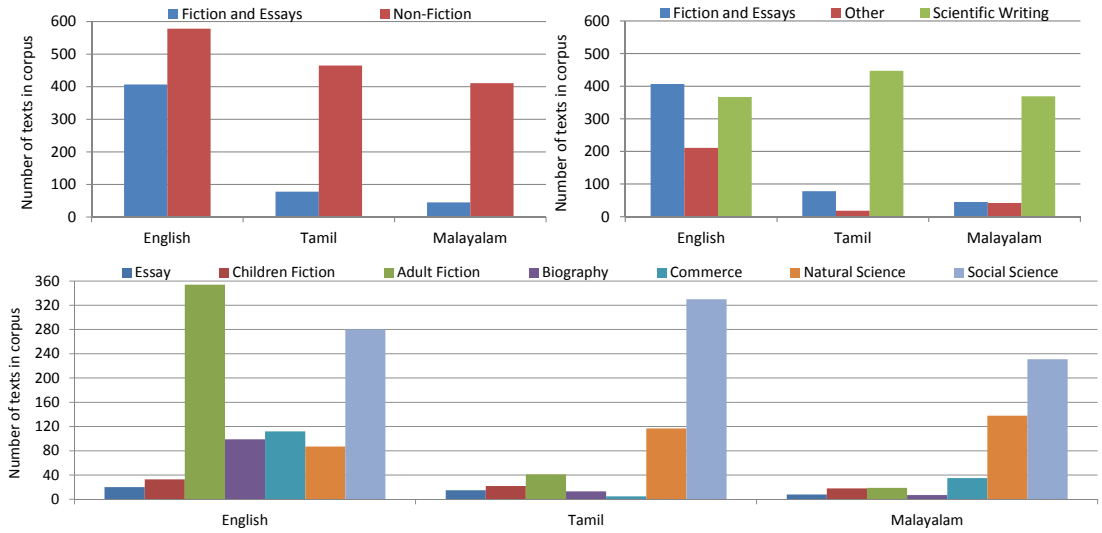

Figure 2.2: Distribution of English texts from the BNC and Tamil and Malayalam texts from the CIIL corpus for classification tasks with different levels of granularity: Two genres (top left), three genres (top right), and seven genres (bottom).

Lee's genre classification of the BNC was taken as a basis to identify texts from the seven genres found in the CIIL corpus. His contribution includes a spreadsheet which allows for straightforward rule based selection of subcorpora, based on genre or other criteria. Table 2.3 shows the rules that were used for classifying BNC texts in this project. The text IDs for each of the seven genres can be found online[5].

The BNC is tokenized by sentence, but not by paragraph. However, since paragraphs are not annotated or easily automatically identifiable in the CIIL corpus either, paragraph based features could not be used anyway. Therefore, no pre-processing was necessary.

Note that the genre distributions in the BNC are very different to those in the CIIL. They also differ considerably from language to language within the CIIL. Figure 2.2 shows the number of texts in each genre for English, Tamil, and Malayalam texts for each of the three levels of granularity.

2.5 Part of Speech tagging

Some of the baselines and proposed approaches to CLGC discussed in this dissertation use features based on PoS tags. Therefore, some of the corpora were automatically annotated by a supervised tagger. Specifically, PoS tags were required for experiments

[5]http://homepages.inf.ed.ac.uk/s0895822/dissertation

with texts from the BC, LCMC, SUC, NYTAC, and TüPP-D/Z corpora. To this end, the *Stanford Log-linear Part-Of-Speech Tagger* (Toutanova et al. 2003) was used. It includes models for English, German, and Chinese texts, trained on the Wall Street Journal (PTB tag set), Negra (STTS tag set), and Penn Chinese Treebank corpora respectively. This covers four of the five corpora mentioned. In order to tag the Swedish SUC texts, the Stanford tagger was trained on data from the 200,000 token Talbanken05 subcorpus, which has been made available through the CoNLL 2006 shared task.

As all cross-lingual PoS-based features use a mapping to the universal and language-independent tag set proposed by Petrov et al. (2012, see Section 3.3), an alternative assignment method could have been used. It would have been possible to map the tags of the training data and train the tagger on the language-independent set. Since this set includes fewer unique tags, it may have improved tagging accuracy, which in turn may have benefited the CLGC methods. However, since PoS tags are also used for target language specific features, where a larger, more fine-grained PoS tag set is desirable, this would have meant to train two separate taggers, and apply both on each text. Furthermore, in practice, a trained tagging model may be available, but not the data it was trained on. For these reasons, it was decided to PoS annotate all texts with a language-specific tag set and map the results to the language-independent set. The universal PoS tag mappings of Petrov et al. (2012) used were *en-ptb*, *de-negra*, *zh-ctb6*, and *sv-talbanken* for English, German, Chinese, and Swedish texts respectively.

Note that the BC, LCMC, and SUC corpora include PoS tags, which have been manually annotated or at least manually verified. These were not used during this project, either directly or as training labels for a supervised tagger. None of the other corpora discussed in this chapter include PoS tags.

Chapter 3

Text Features for Cross-lingual Genre Classification

As mentioned in Section 1.3, there is very little work on cross-lingual text classification that uses features other than word frequencies. These lexical cues are an intuitive choice, since word frequencies (if carefully selected) are known to correlate with topic, sentiment, genre, and other possible target variables. However, since different languages have mostly disjoint vocabularies, they require some sort of mapping, which is typically achieved with either machine translation or a multi-lingual dimensionality reduction technique. On the other hand, there are many different types of features that are being used in mono-lingual genre classification, as already discussed in Section 1.2. Many of these are easily extracted from texts in different languages without the need for linguistic tools (taggers, parsers, etc.) or resources (e.g. parallel corpora). Such features can help to bridge the language gap, since they are comparable across languages. That is, they correlate with style, function, and communicative purpose in similar ways in different languages. Other cross-lingual features can be derived by exploiting PoS taggers, where available, and mapping them from the source to the target language. Furthermore, genre-revealing features which cannot be used across languages, can be exploited by semi-supervised classifiers if additional target language texts are available. In this chapter, I introduce the features used in the experiments of this project. The genre-specific distributions of feature values in different languages are analysed and the strategy for mapping absolute values is discussed.

The features that can be exploited for a classification task depend on the language pair and the available resources. Some languages, for example, have similar punctuation conventions, which makes punctuation frequencies possible as cross-lingual predictors.

In this project, features are therefore grouped by the level of required language similarity and/or necessity for linguistic resources. While this is a loose, and somewhat subjective, criterion, it helps to establish how classification algorithms perform in different scenarios. Sections 3.1 to 3.4 describe these groups of features in detail. Section 3.5 explains how feature values can be scaled in order to adjust for language differences. This standardization is used throughout this project and the resulting relative feature values were used in the figures of this Chapter.

3.1 Text Statistics

Text statistics have been exploited for genre classification ever since research in the area started. Karlgren and Cutting (1994), for example, used features such as average sentence lengths, type/token ratios, and the frequency of long words (>6 characters), among others, to classify genres. Similarly, Kessler et al. (1997) used ratios of word, sentence, character, and word type frequencies, as well as standard deviations. Such features are easily extractable from plain texts without the need for linguistic resources and have been shown to be good predictors in mono-lingual experiments. The findings of Petrenz (2009) show that feature sets including text statistics make very robust genre classifiers in the face of changing genre-topic correlations.

Text statistics lend themselves as cross-lingual features, since they can be extracted in any language, as long as word, sentence, and (preferably) paragraph boundaries can be determined. They do not require translation, parsing, tagging, or other linguistic tools or resources, and are therefore suitable even for tasks including low-resourced languages. Furthermore, their simplicity means that they are likely to correlate with comparable text characteristics across different languages. A high average word length, for example, is likely to be a cue for a more complex text, while a low value could hint at a young target audience. If a genre is defined by such characteristics in a similar way in the source and in the target language, then these statistics should be good cross-lingual predictors.

Many of the text statistics exploited in the experiments for this project have been used in mono-lingual tasks before. However, others are new and go beyond the simple ratios and standard deviations found in previous literature. Since there is no exhaustive list of text statistics, these features were included based on intuition and it is hoped that they inspire a search for further genre-revealing variables for both mono-lingual and cross-lingual tasks in the future. The following list briefly describes how each feature

was extracted from the texts, and provides a motivation for including the newly added predictors.

Document Length

The number of words in a text.

Average Word Length

The average number of characters over all words in a text.

Average Sentence Length

The average number of words over all sentences in a text.

Average Paragraph Length

The average number of words over all paragraphs in a text.

Sentence Length Standard Deviation

The standard deviation in number of words per sentence from the mean.

Paragraph Length Standard Deviation

The standard deviation in number of words per paragraph from the mean.

Type/Token Ratio

The number of unique words (types) divided by the number of all words (tokens). Note that the type/token ratio is strongly affected by the length of a text, as fewer and fewer formerly unseen words will appear as a text grows longer. This is known as Heaps' Law (Herdan 1960). In order to reduce this impact, the type/token ratio is recorded for 300 words, rather than the whole text. That is, the final feature value is the average of the type/token ratio within a sliding window of 300 words. If a text has less than 300 words, its type/token ratio is scaled based on the average of all texts of the same language for the length of the text. That is, if a text of $n < 300$ words has a type/token ratio of x and the corpus averages for sliding windows of n words and 300 words are y and z respectively, the feature value for this text would be $\frac{x*z}{y}$. The threshold of 300 words was chosen so that a large enough number of texts in all used data sets exceed it, which guarantees reasonable averages for scaling.

Number/Token Ratio

The number of numbers divided by the number of tokens. Note that only Arabic numerals are counted, while words designating numbers (e.g. *twenty*) as well as other numeral systems (e.g. Chinese) are ignored.

Single Sentence Paragraph (SSP) Count

The number of paragraphs with only one sentence. This has not been used in genre classification experiments before. The motivation for adding this feature, as

well as the following two, is that single sentence paragraphs are often structurally important parts of a text, such as (sub-)headlines, dates, signatures, or lists. The existence and location of such building blocks within a text might hint at its genre.

SSP/Sentence Ratio

The same as above, but divided by the total number of sentences.

SSP Distribution

This feature indicates the location of SSPs within a text. More precisely, it is the average of the relative distances from the center of the text over all SSPs. That is, the text-initial set of headlines, bylines, and datelines one might expect to find in a newspaper editorial would result in a high value, as would the signature of a letter at the end. A bullet point list of company revenues, on the other hand, would yield a low SSP distribution value, unless preceded or followed by several paragraphs of text.

TF-iDF Average Precision

This feature, as well as all features below, uses term frequencies (TF) and inverted document frequencies (iDF). The TF-iDF score of a word is often used in information retrieval applications to estimate its importance for a text. The intuition is that overall rare words, which feature heavily within a text, are likely to be good indicators of its topic. There a different ways to obtain TF-iDF scores, but for this project, they were computed as

$$TF\text{-}iDF(w,t,C) = TF(w,t) \times \ln\left(\frac{|C|}{DF(t,C)}\right)$$

where $TF(w,t)$ is the raw count of times a word w occurs in a text t and $DF(w,C)$ is the count of texts in a corpus C that contain the word w.

For this feature, the top 10 TF-iDF rated words of a text are identified and their positions within that text are quantified. The measure used here is average precision, which is also commonly used in information retrieval to compare rankings. All occurrences of these top 10 TF-iDF words are marked as *ones*, while all other words are marked as *zeros*. The text is then treated as a ranked list of *ones* and *zeros* and the average precision score is computed as

$$\frac{\sum_{k=1}^{N}(P(k) \times W_k)}{\sum_{k=1}^{N} W_k}$$

where $P(k)$ is the precision at the k^{th} word (i.e. the number of ones in ranks 1 to k, divided by k) and W_k is either one or zero, depending on whether the k^{th} word is among the top 10 TF-iDF words or not. A high feature value means that important (topical) words are clustered around the beginning of a text, while a low value indicates the opposite. The reasoning behind this feature is that one application of genre classification is text summarization. Prior work in that field (e.g. Goldstein et al. 2007; Yatsko et al. 2010) has found that the genre of a text dictates the structure and location of important information within that text. At least for some genres, these conventions hold across language boundaries too. For example, Thomson et al. (2008) show that the inverted pyramid structure in newspaper reportage texts is found in various cultures and languages. The metric of average precision has been chosen to reflect such a pattern, as it is strongly biased towards the first few positions in the ranking. While the average precision of highly ranked TF-iDF words is unlikely to fully capture structural conventions such as the inverted pyramid, it is an attempt to benefit from them and inspire research on similar features for this purpose. Note that the actual words behind the TF-iDF scores, their semantic meaning, or the associated topic(s) are irrelevant and being ignored by this measure, as they are likely to differ from language to language.

Top 10 TF-iDF Scores

The relative TF-iDF scores of the ten words, which are marked as *ones* above, are added as separate features. That is, the absolute scores are scaled so they add up to 1. The combined set of ten features reveals whether a text contains very few high-scoring TF-iDF words or a more even distribution of scores. This is hoped to capture how focussed a text is on a single topic or, conversely, to what extent a text covers multiple subjects.

Average iDF Score

This is the average iDF score over all the words in a text. Note that the term frequency (TF) is not computed. However, words that occur frequently in a text will contribute more to the average, as each occurrence is treated separately. The feature value is therefore equivalent to the sum of TF-iDF scores of each *type* (unique word), divided by the number of *tokens* (document length). The idea behind this feature is to estimate how much the vocabulary of a text differs from that of others in the corpus. A low value is an indication of a high frequency of common function words, which could be a cue for the simpler language found in

children's literature, for example.

Average iDF Score (10%)

This is the same score as before, but the average is computed only for words that are found in less than 10% of the texts in a corpus. That is, common function words such as *the* or *of* in English are ignored. A high feature value can mean that a text contains many uncommon content words, which can be an indication for jargon found in texts targeted at a specific group, such as academic prose.

This list has potential to be extended in future work. The field of information retrieval, in particular, provides many ideas for quantifying certain aspects of a text. One example is clarity (Cronen-Townsend and Croft 2002), which was originally proposed as a measure of a search query's ambiguity with respect to a collection of documents. Applied to a text within a corpus, a clarity feature might allow inference about how unique the vocabulary of a text is when compared to others of the same language. This is somewhat similar to the average iDF score mentioned above. However Cronen-Townsend and Croft (2002) showed that the words contributing strongly to the clarity score are different from the high-ranking iDF words. Text clarity as a feature value might therefore be a valuable addition and is one of many promising leads for future work.

Figure 3.1 shows the relative (see Section 3.5) value range for nine of the above features, broken down by genre in English, Chinese, and Swedish texts. For presentational reasons, the 4-genre categorisation is used here. The boxplots show that at least some of the features have predictive powers for a cross-lingual task. That is, the range of values differs for different genres in similar ways across languages. For example, in all three languages, non-fiction texts tend to have longer words and sentences than fiction texts, as well as more numerals. Similarly, press texts are characterized by a high type/token ratio, short paragraphs, and an average word length longer than fiction, but shorter than non-fiction.

Not all of the features separate the genres in the same way across languages, at least for the shown corpora and level of genre-granularity. For example, while press texts have relatively many paragraphs with only one sentence in English and Swedish, this is not the case for Chinese texts. Similarly, they seem to have relatively few numbers written in Arabic numerals in Chinese and Swedish, while this is not the case in English. Note that this does not mean that such features are no good predictors of genre within a language. The distribution of SSP ratios in press texts, for example, differs from that of fiction texts in all three languages. These correlations mean that the feature might be well suited as a target language specific feature. That is, while it may be of little use to

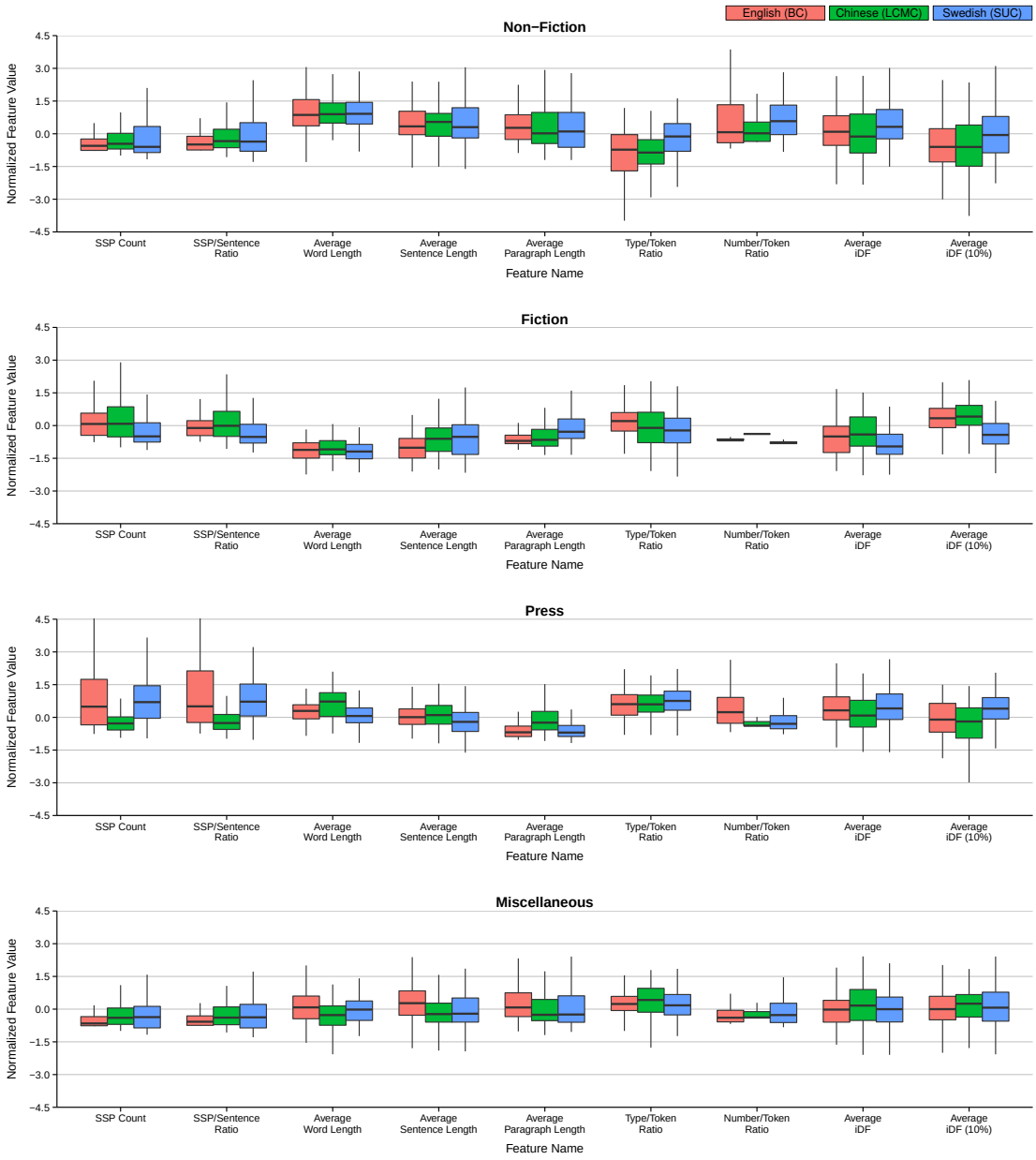

Figure 3.1: Boxplots of the standardized (see Section 3.5) values of nine text statistics features. The black horizontal line depicts the median. The lower and upper box edges represent the first and the third quartile respectively. The whiskers show the extent of variability beyond these quartiles. The texts come from the BC (English; red boxes), the LCMC (Chinese; green boxes), and SUC (Swedish; blue boxes), separated into four genres (see Figure 2.1).

bridge the language gap, it may help to separate genres in the target language after this gap is bridged, thus improve classifier accuracy.

3.2 Punctuation Marks

Like text statistics, features derived from punctuation marks have been used to classify genres before (e.g. Lim et al. 2005; Finn and Kushmerick 2006). Stamatatos et al. (2000a), for example, empirically show that adding the frequencies of common punctuation marks to a set of word based features increases classification accuracy and reliability, especially where a limited amount of training data is available. While both conventions concerning punctuation and the actual marks used can vary from language to language, there are also similarities. Question marks, for example, have identical functions in many languages. It is therefore possible that punctuation marks can be exploited as cross-lingual genre predictors, at least for some language pairs.

I used the frequencies of 32 punctuation marks as features, both for cross-lingual classification and for subsequent improvements within the target language. While somewhat more sophisticated features have been derived from punctuation for monolingual genre classification tasks before (e.g. the frequency of *, where* in English texts (Kessler et al. 1997)), simple punctuation one-gram frequencies were assumed to be most promising in a cross-lingual setting.

Due to the difference in punctuation symbols of Chinese texts, compared to English and Swedish ones, such features were not used in the experiments with data from the BC, LCMC, and SUC. Punctuation frequencies were however used as cross-lingual predictors for newspaper sub-genres in English and German, as well as in the experiments with the Reuters, JRC-Acquis and Europarl corpora. Figure 3.2 shows the value range for five of the above features, broken down by genre in English and German: The relative frequencies of question marks, parentheses, periods, commas, and hyphens. They were selected both because they are relatively common, and because they show how some punctuation mark features correlate with genre more than others. Note that each graph in Figure 3.2 represents a different feature, rather than a different genre as in Figure 3.1. This is due to presentational reasons, as fewer features, but more genres are used for visualization here.

It can be observed that interviews in the *New York Times* have a high frequency of question marks, compared to other texts in the same paper, especially reportage. This is intuitive and has been suggested for English texts before (e.g. Stamatatos et al. 2000a).

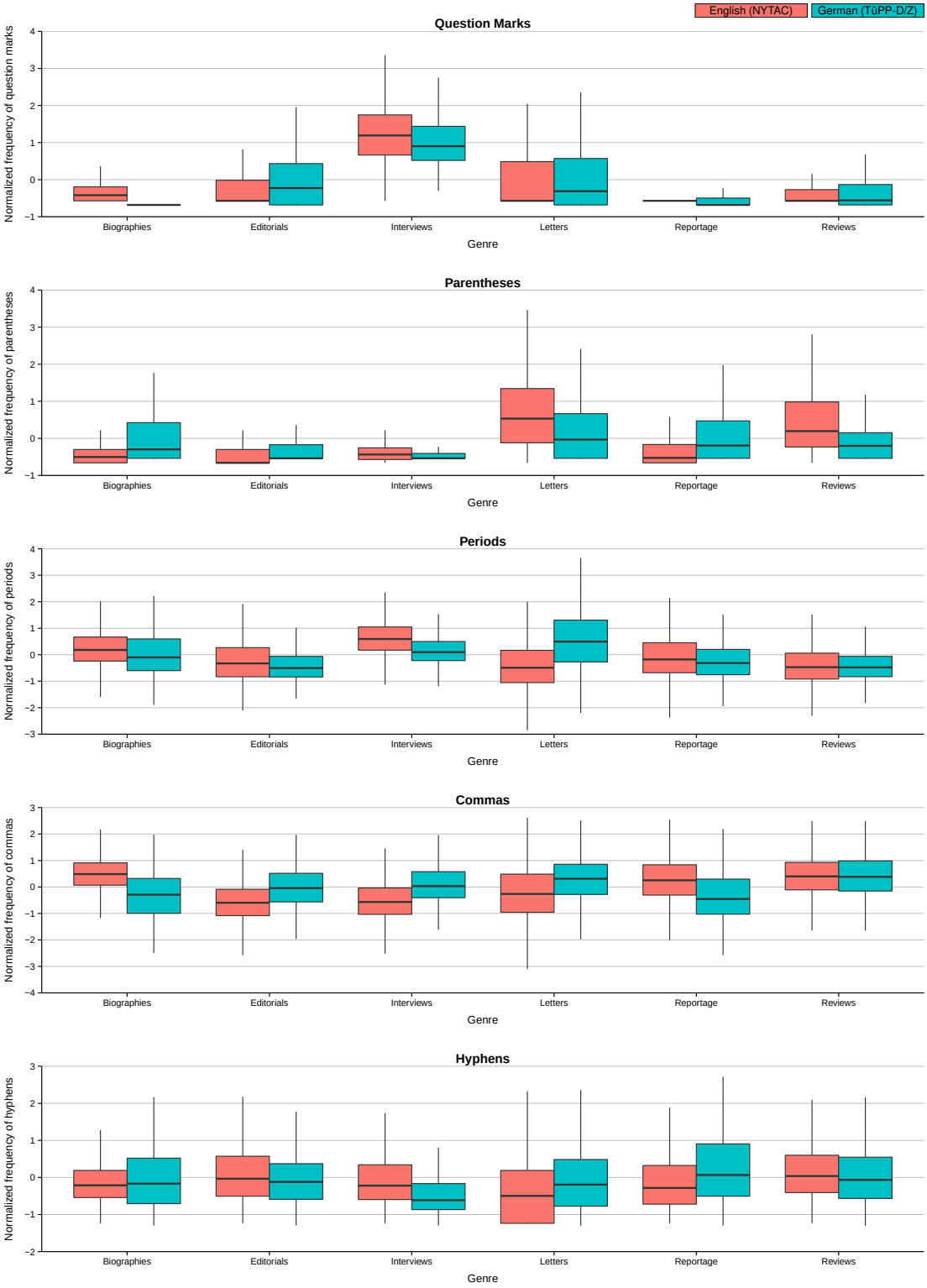

Figure 3.2: Boxplots of the standardized (see Section 3.5) frequencies of five punctuation marks. The black horizontal line depicts the median. The lower and upper box edges represent the first and the third quartile respectively. The whiskers show the extent of variability beyond these quartiles. The texts come from the NYTAC (English; red boxes) and the TüPP-D/Z (German; blue boxes), separated into six genres.

The fact that it is also true for the German newspaper *tageszeitung*, means that question mark frequencies can be used as a relatively reliable feature to distinguish interviews across this language boundary. A similar, albeit less clear, correlation can be observed for parentheses, as they seem to occur relatively frequently in letters to the editor written in either language. The frequency of periods also has similar genre-specific distributions, as they seem to be more frequent in interviews than in reviews or editorials, for example. However, their frequency in letters differs strongly between English and German texts. Commas, while probably reasonably good predictors within either English or German texts, have different genre-specific distributions across languages. The exception are reviews, which contain relatively many commas in both the NYTAC and the TüPP-D/Z. Lastly, the distribution of hyphen frequencies seems to be affected by the genre only marginally, in either language.

Another interesting observation that the plots in Figure 3.2 allow is that letters tend to have a higher variance of punctuation frequencies than other genres. This is true for most features in both languages. One explanation could be that letters to the editor, unlike all other shown genres, are typically written by readers, rather than journalists. This means that they cover a wider range of styles, as they are not bound to linguistic rules laid out in newspaper style manuals.

3.3 Part of Speech tags

Most research on mono-lingual genre classification has exploited PoS taggers. While the early work of Karlgren and Cutting (1994) relies on hand-picked PoS tag frequencies, such as adverbs and prepositions, later approaches have often used PoS trigrams (e.g. Argamon et al. 1998; Sharoff 2007) or PoS histograms (Feldman et al. 2009, see below). One of the few genre classification studies on non-English texts (German in this case) exploits similar features (Wolters and Kirsten 1999). One of the reasons for their popularity is that PoS tags can reveal the structure of a text, without including semantic information. This means that text genres can be classified more independently of text topics by using PoS frequencies, than word frequencies, for example. Petrenz and Webber (2011) show this to be mostly true, although some PoS tags (common nouns in their experiments) are affected by topic also.

As PoS tags contain no semantic meaning, they potentially could be valuable cross-lingual predictors for the genre of a text, as the correlation between topics and genres is likely to differ in the source and target languages. However, there are two problems

when using PoS tags for this task. Firstly, texts in both languages need to be tagged automatically, which requires training data manually annotated with PoS information. This would typically be available in the source language, seeing as genre-annotated text is assumed to exist. However, it may not be available for the target language. Secondly, PoS tag sets differ from language to language. This is both because of differences in grammars (e.g. case) and because of differences in granularity of the tag sets used, which can vary even for different sets in the same language. These variations mean that PoS tag frequencies cannot be meaningfully used as predictors across languages without some form of adaptation.

For the experiments of this project, I use the universal PoS tag set proposed by Petrov et al. (2012) to remedy this problem. This resource has not been used in work on genre classification before. Petrov et al. (2012) have mapped 30 PoS tags for 25 languages to a universal set of 12 PoS tags. These are nouns, verbs, adjectives, adverbs, pronouns, adpositions (pre- and postpositions), determiners, numerals, conjunctions, particles, punctuation marks, and *others*. For instance, the RB, RBR, RBS, and WRB tags of the Penn TreeBank are mapped to the adverb category. I use universal PoS frequencies to bridge the language gap in some of the experiments of this project. More complex PoS based features, which reveal deeper structural properties of a text, were not exploited, as grammar differences are likely to affect the predictive power of such features. However, it is possible that much can be gained in CLGC from some of these features, especially for language pairs with similar grammar rules. Further research into such cross-lingual features and their benefits would be an interesting topic for future projects.

Mapping from a set of dozens or even hundreds of tags to a small set of twelve comes at a cost. For example, the *verb* category of Petrov et al. (2012) does not reveal whether a text is written in present or past tense – unlike the more fine-grained categories in the English Penn TreeBank and other tag sets. Such information, however, is highly useful when classifying genre, at least for English texts (Petrenz and Webber 2011). While it is necessary to sacrifice this high level of granularity for using PoS tags as predictors across languages, it can be retained for target language specific features. That is, in semi-supervised learning approaches, distances between target language texts can be computed based on the full set of PoS tags in that language, rather than the universal mapping.

For this reason, the PoS histogram feature set by Feldman et al. (2009) was adopted for all PoS based connections between target language texts. This contains the means

and standard deviations for each tag across sliding windows of five words. Formally, Feldman et al. (2009) describe the process as follows:

Let n be the length of a text and let K be the number number of PoS tags in the tag set. For each sliding window $j \in \{1,...,n-4\}$ calculate a histogram $h_j \in R^K$. Let $H = \{h_1,...,h_{n-4}\}$ and let $\mu(H)$ and $\sigma(H)$ be the mean and standard deviation of H, respectively. The feature vector is then $[\mu(H)\sigma(H)]^T$.

This approach captures the variation of identical tags found in close proximity to each other, in addition to their frequencies. It also, unlike the n-gram approach, results in a dense set of $2 \times K$ features. Note that Feldman et al. (2009) used principal component analysis on their feature set. This step is omitted in this project, as preliminary experiments found no significant difference in results and the dimensionality of the feature space is already relatively low.

Figure 3.3 shows the relative (see Section 3.5) value range for eleven of the universal PoS tag frequencies, broken down by genre in English, Chinese, and Swedish texts. The values for determiners are not shown here, as the mapping by Petrov et al. (2012) does not translate any Swedish tag to this universal tag. Note that determiner frequencies are still used for English to Chinese (and vice versa) classification experiments, as well as for the English to German experiments using the NYTAC and the TüPP-D/Z.

It is clear that the distributions of most tags are affected by genre. One can also observe strong similarities for genre-specific distributions across languages. For example, non-fiction is characterized by comparatively many nouns and adpositions, but few verbs, adverbs, pronouns, and punctuation marks in all three languages, while the opposite is true for fiction texts. Press texts, like non-fiction, have relatively many nouns and few verbs and pronouns. However, they do not contain as many adpositions or as few punctuation marks. The *Miscellaneous* category, as already observed in Figure 3.1, is harder to distinguish based on these features, as none of the tags is particularly frequent or infrequent in this genre.

Not all of the tags have similar frequency distributions across all three languages. For example, in English and Swedish texts, numerals are more frequently found in non-fiction than in fiction. This is not true for Chinese, where there is little difference between these two genres. In most, but not all, such cases, the genre-specific distribution is similar in the BC and SUC corpora, but differs in the LCMC.

Overall, the genre-specific distributions of universal PoS tag frequencies are promising. At least some of the tags seem to correlate with genre in similar ways across languages, which would make the respective frequencies good features in a CLGC task.

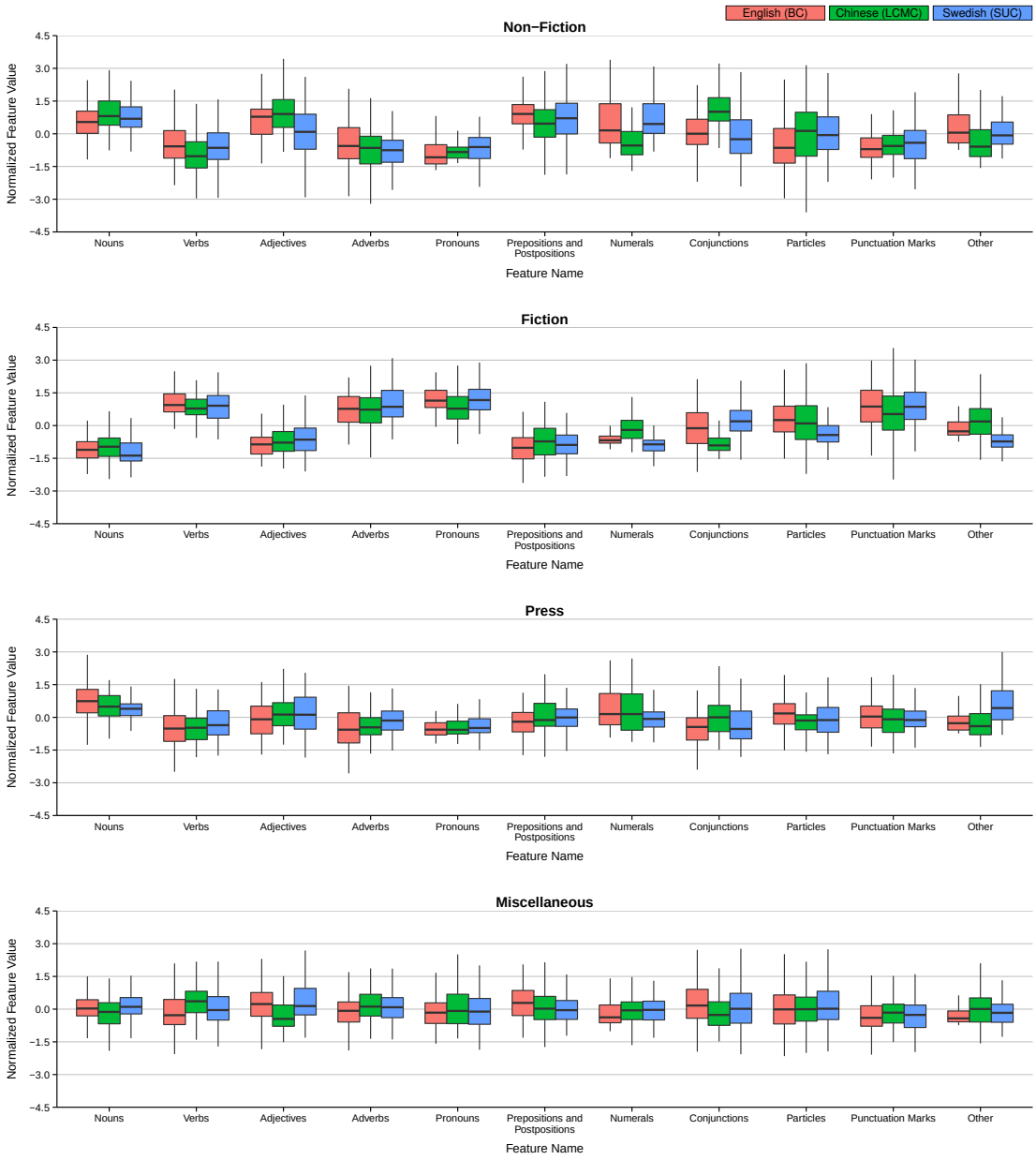

Figure 3.3: Boxplots of the standardized (see Section 3.5) frequencies of eleven universal PoS tags (Petrov et al. 2012). The black horizontal line depicts the median. The lower and upper box edges represent the first and the third quartile respectively. The whiskers show the extent of variability beyond these quartiles. The texts come from the BC (English; red boxes), the LCMC (Chinese; green boxes), and SUC (Swedish; blue boxes), separated into four genres (see Figure 2.1).

3.4 Word Frequencies

Features based on word occurrence are the most intuitive and most common predictors in text classification approaches. Depending on the actual task, there are many possible variations, both in the selection strategy (e.g. all words, only content words, only adverbs, etc.) and the feature values (binary, counts, ratios, TF-iDF scores, etc.). Word-based variables have previously been exploited to predict genre categories. Often, these were frequencies of hand-picked function words, which were chosen to correlate with genre, but not topic. Examples of such words (or word n-grams) used in mono-lingual genre classification in English are *therefore*, *of course*, *shall*, or *hardly* (Karlgren and Cutting 1994; Kessler et al. 1997; Ferizis and Bailey 2006). Alternative methods use more automatically extractable and less language-specific features, such as the 50 most common words (Stamatatos et al. 2000a), all words (Freund et al. 2006), or word n-grams (Sharoff et al. 2010).

As different languages, which are not closely related, have little semantically meaningful vocabulary overlap, word frequencies as such cannot reasonably be used as cross-lingual features. The common remedy is machine translation (MT). As this method requires cross-lingual resources, such as parallel corpora, it is restricted to language pairs, for which such resources exist in sufficient quality and quantity. While none of the approaches in this project relies on the availability of MT, the label propagation method discussed in Chapter 7 can exploit translated word frequencies as additional features to boost performance. Rather than using full text translations (like the baseline in Section 4.1), these frequencies are obtained by translating single words from the source into the target language, or vice versa. While I used the *Google Translate* API for all translation tasks, bi-lingual dictionaries can be used instead of MT systems to map single words.

The downside of the approach is that a limited number of words have to be chosen for translation. One option would be hand-picked words that have been shown to distinguish genres well in English, such as those mentioned above. However, this project is not restricted to English as the source language. Therefore, more language-independent selection methods had to be considered. Another option would be a supervised selection based on the genre labels in the source language. This is employed by Prettenhofer and Stein (2010), for example, as they use information gain for selecting suitable pivot features for their cross-lingual structural correspondence learner. However, these words may be very topical, due to the correlation between genres and topics (see Petrenz and

now	only	years	we	also	its	people	first	most	do
into	after	said	time	you	mr	them	can	two	some
year	other	last	over	just	because	like	no	even	will
could	so	up	were	than	what	how	many	if	there
had	way	out	much	would	m	i	then	where	it's

Table 3.1: 50 words from texts of six genres in the NYTAC corpus, which occur in closest to 50% of all texts.

Webber 2011). Furthermore it allows translation only from source to target language, as no genre labels exist in the target language and therefore candidate words cannot be selected in a supervised fashion. The third option is an unsupervised selection. One option is use the top k words of one or both languages, as Stamatatos et al. (2000a) have done for mono-lingual genre classification. Another possibility is to use words that split the corpus as evenly as possible, that is words that occur in close to 50% of all texts, regardless of genre. The intuition is that such words are too common to carry much semantic meaning (unlike *intergalactic* or *electricity*), but rare enough to distinguish texts based on whether or not they are present (unlike *the* or *of*).

A preliminary experiment showed that the latter set of words outperformed both the most common words and the information gain selection in predicting genres across languages. It was therefore used in the label propagation method (Chapter 7). Table 3.1 shows the top 50 of these words, that is those that occur in roughly half of all texts, based on the NYTAC sub-corpus used for experiments. While this list includes content words, such as *people*, most words are function words, which have traditionally been used for genre classification. Furthermore, the list includes words that reveal subjectivity (*I, we*), tense and/or aspect (*do, will, were, said, can, could*), and formality (*it's*), all of which may be helpful in cross-lingual tasks, provided that it can be appropriately translated and both languages have similar genre conventions in these respects.

In the label propagation experiments described in Chapter 7, the top 1,000 words in the target language are extracted and translated into the source language, using *Google translate*. They are then ranked based on how evenly their occurrence splits the source language corpus. The frequencies of the top 500 word pairs of this ranking are used as cross-lingual features. Furthermore, the top 500 untranslated words of the target language are used to compute distances between target language texts only.

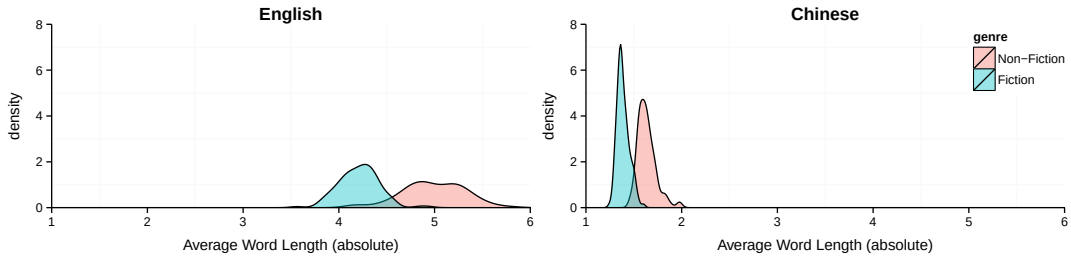

Figure 3.4: Kernel density estimates (Gaussian kernel, bandwidth chosen by *Silverman's rule of thumb*, as implemented in the *R stats* package (R Core Team 2012)) of the average word length in fiction (blue) and non-fiction (red) texts in English (left) and Chinese (right).

3.5 Scaling Feature Values

In most classification problems, data sampling is assumed to be independent and identically distributed (i.i.d.), that is both the data used for training and the data used for testing is sampled from the same probability distribution. In a cross-lingual task however, this would typically not be the case, as the value of a variable is not only affected by the class (in this case, the genre of the corresponding text), but also by the language itself. An example for this is the type/token ratio, which is a proxy for the vocabulary richness of a text. While some genres might use more unique words than others, and while this might hold across different languages, it is also true that texts in morphologically rich languages, such as Finnish, will typically have high type/token ratios compared to English texts. Using absolute feature values without some form of adaptation or mapping to the target language can therefore harm classification results.

To further illustrate the problem, consider the kernel density estimate graphs in Figure 3.4. They show the estimated density function of the average word lengths over all words in a text, based on data from the BC and the LCMC for fiction and non-fiction texts. The *Press* and *Miscellaneous* classes were omitted here for simplicity. The feature separates the two genres reasonably well and similarly in both languages: In both English and Chinese, words in fiction texts tend to be shorter on average than those in non-fiction texts. This makes sense intuitively, considering that the latter class is comprised of academic prose and official reports (see Table 2.1). However, the absolute values are very different in the two languages. Unsurprisingly, Chinese words contain fewer characters on average than English words. Therefore, if a classifier was trained on English texts and established a decision boundary at the intersection of the red and

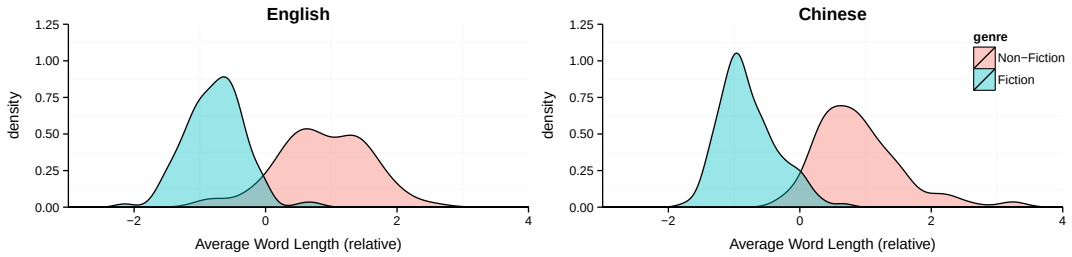

Figure 3.5: Kernel density estimates of the standardized (i.e., relative) average word length values in fiction (blue) and non-fiction (red) texts in English (left) and Chinese (right).

the blue curves, it would classify all Chinese texts as fiction.

In order to reduce the impact of the language, the feature values of source language texts have to be adapted to the target language, or vice versa. In this project, this was done by standardizing cross-lingual feature sets to zero-mean and unit-variance. That is, the value of a given feature for a given text was scaled by subtracting the feature's mean over all texts and dividing the result by the feature's standard deviation:

$$x' = \frac{x - \mu}{\sigma}$$

Note that standardization (or another form of scaling) is often used in machine learning to balance the impact of features among each other. However, typically the scaling factors of the training data are used and applied to the test data as well. While the benefit of balancing feature impacts was appreciated, the aim of standardizing values in this project was to reduce the impact of language. Therefore, scaling was done for each language separately, that is the means and standard deviations of the source language feature set were not used to scale target language features. The process is unsupervised, that is the genres of texts (where known) were ignored when scaling.

Consider the kernel density estimates in Figure 3.5. They show the same density functions for the same data as Figure 3.4, except that the average word length values have been standardized, that is they use relative, rather than absolute, values. While the relationship between fiction (shorter words) and non-fiction (longer words) remains intact, the impact of the language has been greatly reduced. A decision boundary at the intersection of the curves in the left graph (English) would now make a very good classifier for Chinese texts as well.

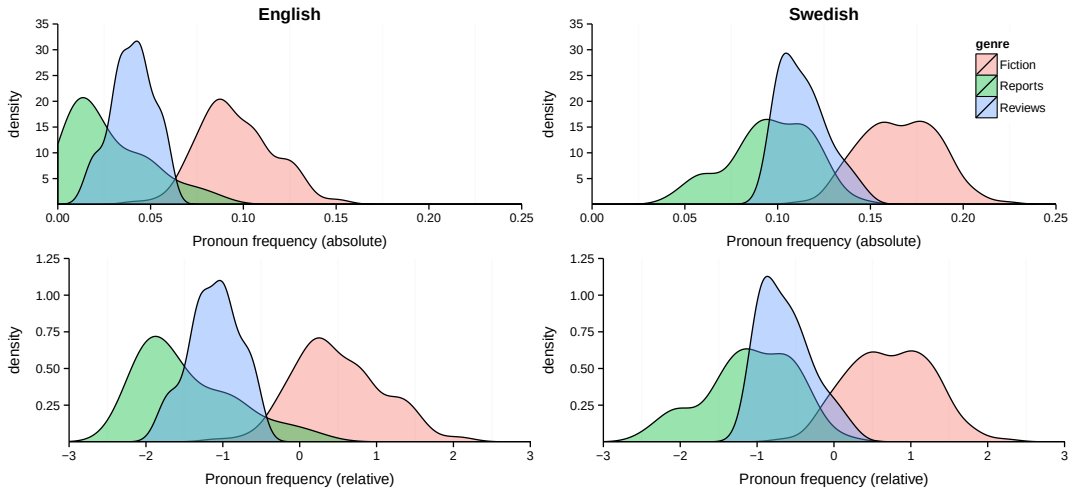

Figure 3.6: Kernel density estimates of the absolute (top) and relative (bottom) pronoun frequencies in fiction texts (red), reports & official documents (green) and newspaper reviews (blue) in English (left) and Swedish (right).

The example of word lengths in English and Chinese texts may be an extreme case of language differences, as it is obvious that the number of characters per word cannot be used as a meaningful feature between these two languages in its raw form. However, the impact of language on the values of the features described in this chapter can be observed in many other, less obvious, cases as well. Figure 3.6 shows the kernel density estimates for raw and standardized frequencies of pronouns in English (BC) and Swedish (SUC) texts for three of the nine genre categories. Again, the feature is a good predictor, as the distributions are different for each genre, but similar across languages. However, this is only the case after scaling, since more words are classified as pronouns in Swedish texts than in English texts by the PoS tagger. In fact, only one of the 483 texts in the BC has more than 15% of words tagged as pronouns, whereas that is the case for almost a quarter of all texts and over 75% of fiction texts in the SUC. Figure 3.6 shows that such differences can be reduced by using relative values, as they use identical means and standard deviations.

While standardizing feature sets separately is an appropriate solution to the language bias problem for the experiments in this project, there are potential issues. Firstly, it requires a sufficiently large set of unlabelled texts in the target language in order to calculate feature means and standard deviations. However, this is assumed to be available for the classification methods presented in Chapters 5 and 7 anyway. Another problem would be a strong difference in the distributions of genres in source and target

language corpora used to compute the scaling factors. Unless a feature's values are not affected by the genre of the texts (in which case it would be a poor feature for this task), its mean and standard deviation are affected by the genre distribution of a text collection. Therefore, such differences would lead to unwanted biases in the scaling factors.

With the exception of the BNC and CIIL corpora (Section 2.4), which are used only for a proof-of-concept experiment in Chapter 7, the text collections used in this project have identical or very similar distributions of genres, and no adjustment was made for this reason. However, since genre labels are not assumed to be known in the target language, their distribution might be unknown as well. An alternative way to obtain scaling factors in this case could be the means and standard deviations from a separate set of texts (i.e. not involved in the actual classification task) of a known genre in both the source and target language. This could, for example, come from *Wikipedia* texts, which are available in a wide range of languages. The resulting feature values would then be relative to that genre in both languages. Such experiments are left to future work.

One further restriction is that the standardization method assumes numerical feature values to make sense. While binary features (such as presence vs. absence of a word) have been considered during the early stages of this project, none of the results described in this thesis make use of binary values. Note that this does not restrict the choice of continuous (e.g. type/token ratio) or discrete (e.g. document length) variables and both types of features are used in this project, although the values after standardization will always be continuous. All experimental results described in Chapters 5, 6, and 7, as well as those for the TSVM baseline (Section 4.4) were obtained by standardizing feature values as outlined in this section.

Chapter 4

Baselines

Experimental results are meaningful only in comparison to a baseline that uses the same data. Unfortunately, there are no published results on the data sets used in this project. This is due to the fact that there is no prior work on CLGC, except for the approach of (Snyman et al. 2012), which is not feasible for anything but very closely related languages (see Section 1.2). Therefore, all the baseline performances had to be evaluated as part of this project. To this end, four approaches were chosen: A simple full text machine translation method, two formerly proposed cross-lingual techniques, and an out-of-the-box transductive SVM classifier. This chapter provides an overview and experimental results for the different baselines against which new results will be compared.

4.1 Full Text Machine Translation

One of the most intuitive ways to approach any cross-lingual NLP task is to use translation, either manual or automatic. For the task of cross-lingual text classification, this may mean translating entire texts from the source language into the target language, or vice versa. Afterwards, the problem can be treated as a mono-lingual task, for which previously proposed machine learning techniques and feature sets are likely to exist.

For this project, two versions of a simple and intuitive, though resource-intensive, baseline were implemented to evaluate such an approach. The first uses the *Google Translate* API to translate all genre-annotated source language texts into the target language. Subsequently, a mono-lingual classifier is trained on the translated source language texts and tested on the original target language texts. The second version translates all unlabeled target language texts into the source language. A classifier is

then trained on the untranslated source language text and tested on the translated set. The feature set is the same for all language pairs and translation directions. To keep this baseline simple, word frequencies as proposed for genre classification by Freund et al. (2006) were chosen. As the feature set can be automatically created from a corpus, its assembly does not depend on genre-specific knowledge in the classifier language, unlike the hand-picked predictors of Karlgren and Cutting (1994), for example. All translations were carried out between the 19th and the 21st of September 2013. Since the output of *Google Translate* for translations into Chinese is not tokenized by words, overlapping character 2-grams were used as features in these cases. While mistakes introduced by the machine translation system are likely to affect classification performance, the high costs of manual translation would reduce the usefulness of an automatic cross-lingual genre classification method. Manual translation was therefore not considered as an option for this baseline, or any other part of this project.

Note, however, that the use of machine translation system comes at a cost too. As mentioned in Section 1.3, any approach relying on the availability of cross-lingual resources, such as the massive parallel corpora needed to train effective MT systems, is restricted to relatively well-resourced languages, which is why even *Google Translate* uses English as a pivot for translating between arbitrary language pairs (Boitet et al. 2010). Since the aim of cross-lingual classification is to remedy a lack of resources (genre labels in this case) in the target language, such dependence may be problematic. Furthermore, it adds computational overhead, in particular if complete texts are translated. Indeed, for online applications, the cost of translating every new text into the source language may be prohibitive. In these cases, one would be left with the option of translating the set of source language texts into the target language before training the classification model. Unfortunately, this approach also has disadvantages, as good genre-revealing features are more likely to be known for the source language than the target language.

While this baseline might be simple and intuitive, it cannot be expected to be outperformed easily. Prettenhofer and Stein (2010), for example, use a full text target to source translation baseline in a sentiment classification task and report better performance of their own method only for some target languages and text domains. To evaluate baseline performance, an SVM classifier was trained and used to predict genre labels for the target language texts. SVMs are a popular choice for text classification, as they can handle high-dimensional feature sets efficiently. They have been used for genre classification by Kim and Ross (2008), Sharoff et al. (2010), and others. For all SVM

experiments of this project, unless otherwise stated, I used the implementation in the *R e1071* library (Meyer et al. 2012; Chang and Lin 2011) and its standard parameter values: A radial basis function kernel with the γ parameter, which determines how wide or narrow the kernel is, set to $\gamma = \frac{1}{F}$, where F is the number of features. The C constant of the regularization term, which determines the smoothness of the decision boundary, was set to $C = 1$. Results are presented and discussed in Section 4.5.

4.2 Multi-Lingual Domain Models

Gliozzo and Strapparava (2006) propose a cross-lingual method that exploits words, which are identical in two languages of a comparable corpus. These are often names or abbreviations, such as *Obama* or *AIDS*, but Gliozzo and Strapparava also mention different examples, such as the word *virus*, which is identical in English and Italian. The idea behind their approach is that identical words in both languages carry identical meaning. Their occurrences in texts can be exploited to create semantic links between other, language-specific words in both languages through their correlation with these words. In their topic classification experiments with English and Italian texts, Gliozzo and Strapparava (2006) focus on nouns, verbs, adjectives, and adverbs, to keep semantically unrelated identical words to a minimum.

To build their classifier, Gliozzo and Strapparava first perform Latent Semantic Analysis (LSA, Deerwester et al. 1990) on a document-word matrix which includes texts from both source and target languages. They then use the results to construct a multi-lingual domain matrix, which projects feature spaces from both languages into a common, low-dimensional space. This is then used to train and test an SVM classification model. For a formal description of their method, see (Gliozzo and Strapparava 2006). The authors report good cross-lingual results in their experiments, where they work with newspaper texts covering four different topics. However, the baseline they outperform is weak: A classifier that uses word frequencies directly, without any form of transformation or translation.

In this project, the multi-lingual domain model approach was re-implemented as described in (Gliozzo and Strapparava 2006), as it is one of the few cross-lingual approaches that does not rely on the availability of machine translation systems or other cross-lingual resources. However, the approach may be less suitable for the problem of CLGC in general, and for certain language pairs in particular. Some of the reasons can be understood from the graphs in Figure 4.1. They show the ten words which have the

highest product of document frequencies in both languages, for the English-Swedish and the English-Chinese task. These are the words that the approach of (Gliozzo and Strapparava 2006) relies on to establish the projection to the multi-lingual space. Firstly, it is obvious that this is problematic in the English-Chinese case, as the two languages use different writing systems and overlap is therefore extremely sparse. The top ten overlap words are almost exclusively single letters, which are unlikely to correlate with genre or genre-revealing words in either language.

For English and Swedish more interesting common words exist, some of which have a strong representation in both the BC and the SUC. The most common ones are, however, different from the proper nouns that Gliozzo and Strapparava (2006) reason with. This possibly stems from the fact that texts from a single genre (newspaper text) was used for their experiments, while here a variety of topics, genres, and publication dates are represented. Furthermore, English and Swedish are both Germanic languages, while Italian is a Romance language. This language family difference may have helped in restricting common words to (mostly) proper nouns. This in turn may have helped the English-Italian classifier performance, as some of the English-Swedish common words from Figure 4.1 differ semantically. More precisely, the Swedish words *far*, *be*, and *god* translate to *father*, *ask*, and *good* respectively in English. Also, the words shown in Figure 4.1 are intuitively are more topic-revealing than genre-revealing.

The classifier performance was evaluated on the data from the BC, LCMC, and SUC corpora. Results, comparison, and a discussion can be found in Section 4.5.

4.3 Structural Correspondence Learning

Another cross-lingual classification method was proposed by Prettenhofer and Stein (2010). They use a multi-lingual extension of the Structural Correspondence Learning (SCL) algorithm, which was originally published by Blitzer et al. (2006) as a (mono-lingual) domain adaptation solution. Prettenhofer and Stein evaluate their method for a sentiment classification task with two classes (positive and negative reviews), using texts written in English (source language), German, French, and Japanese. They do however imply that the method works for other problems, such as spam filtering or topic classification.

The authors first exploit a set of labelled texts in the source language to identify k words with the highest information gain with respect to the target variable (i.e. positive or negative sentiment). These words are then translated into the target language one

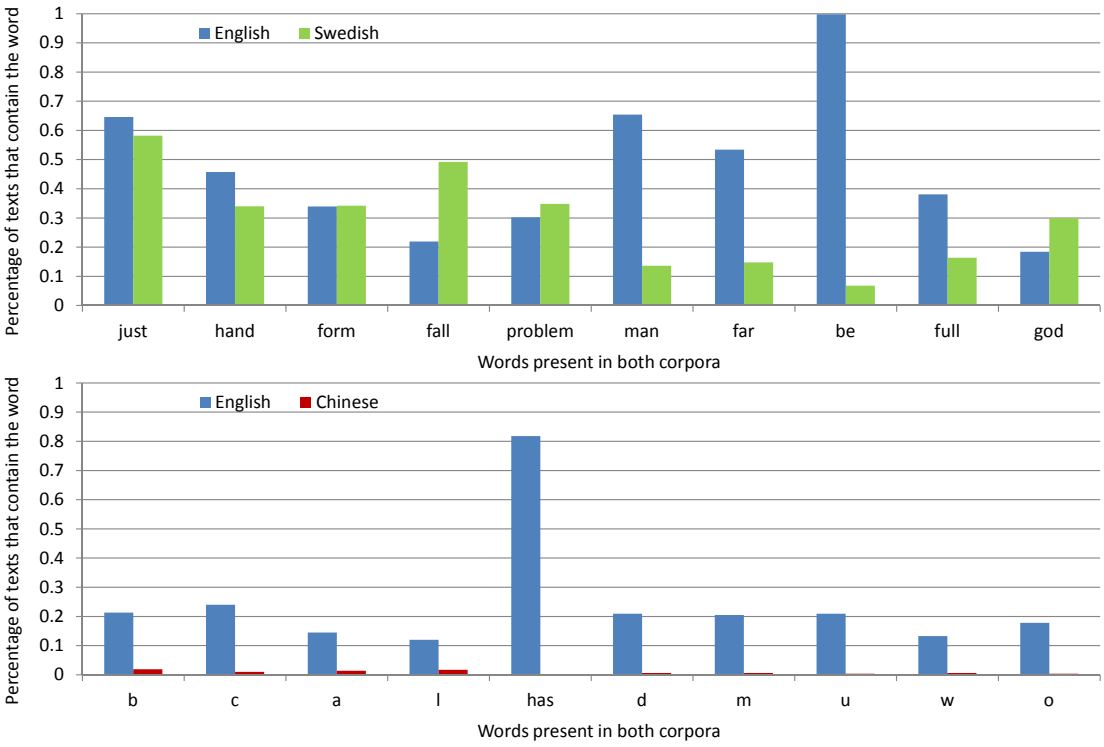

Figure 4.1: Top ten words (nouns, verbs, adjectives, and adverbs only) based on the product of corpus fractions which contain the word. Top: English (BC) and Swedish (SUC). Bottom: English (BC) and Chinese (LCMC).

by one, that is out of context. Prettenhofer and Stein (2010) use *Google Translate* for this task. They then keep only the $k' \leq k$ pairs of words, where each word in a pair is found frequently in its respective language. These k' word pairs are called *pivots* and are subsequently used as target variables for a set of k' linear classifiers. The feature sets for these classifiers include all words in both languages, with the exception of the pivot pair to be predicted. Vocabulary overlap is ignored, that is identical words in the two languages are treated as distinct features. A collection of texts from both source and target languages are then used for training. The idea behind this approach is that the learned weight vectors of these linear classifiers capture the correlations of pivot pair words and other words of both languages. Lastly, Prettenhofer and Stein (2010) find the correlations across pivots by computing the SVD of a matrix which contains the aforementioned k' weight vectors. They then use the top columns of the resulting U matrix as a multi-lingual projection, which can be used to transform the original word frequency based representations of either language into a common form. This allows a classifier to be trained and tested using the projected features. For a formal description

didn't	went	looked	you	said	knew	eyes	him
me	got	saw	don't	her	she	back	my
turned	took	came	your	he	stood	asked	seemed
told	door	thought	night	get	man	walked	head
oh	smiled	has	gone	face	president	program	go

Table 4.1: 40 words with the highest information gain in the BC.

of their algorithm, (see Prettenhofer and Stein 2010).

In this project, cross-lingual SCL is used as a baseline for CLGC. The algorithm was implemented as described by the authors, including choices for thresholds. It differed in adjustments due to the multi-class task of this project, as opposed to the two-class problem assumed by Prettenhofer and Stein (2010). Table 4.1 shows the 40 words most highly ranked by information gain (see Section 6.1 for a formula), with respect to the nine genres in the BC. A strong presence of past tense verbs, as well as pronouns can be observed and such parts of speech are intuitively useful in genre classification tasks. These identified pivot candidates are, predictably, very different from those reported by Prettenhofer and Stein for their sentiment classification task (e.g. *beautiful* and *boring*). However, this fact might pose problems in the single word translation step. The correct translation of the word *you* as a pronoun in German, for example, can be *du*, *dir*, *dich*, *ihr*, *euch*, *Sie*, *Ihnen*, or *man*, depending on grammatical case, number, and formality, none of which can be established without context. This can be true for content words, such as *beautiful* as well. However, in that case, all the German forms (e.g. *schön*, *schöne*, *schönes* etc.), convey the same positive polarity for a sentiment classification task, so any of these choices could be considered helpful. This is less obvious for different translations of function words for a genre classification problem. For example, *didn't* was translated by *Google Translate* as *inte* in Swedish, which is an adverb and simply means *not*. This clearly has a much less specific meaning than *didn't* and does not contain information about tense.

Despite such reservations, the algorithm was evaluated and compared to the other baselines. Results are presented and discussed in Section 4.5.

4.4 Transductive SVM

One of the classifiers proposed in this thesis is a multi-layer label propagation method (see Chapter 7). Since this graph-based algorithm is inherently transductive (that is, all texts to be classified need to be available at training time), it is appropriate to use a baseline with the same restrictions. Transductive Support Vector Machines (TSVM) use both labelled and unlabelled data to maximize the margins of a hyperplane which separates the classes. They have been proposed as particularly well-performing in text classification problems where large quantities of unlabelled data is available (Joachims 1999). For the experiments of this project, the TSVM implementation in the SVMlin library (Sindhwani and Keerthi 2006) is used. SVMlin provides a choice of different inductive and transductive algorithms. For this project, algorithm 2 (Multi-switch Transductive SVM) was used, while the default values were kept for all other options of the library.

To compare TSVMs with the label propagation method, both exploit the same cross-lingual features. The results shown in Section 4.5 were achieved using a combined set of text statistics (Section 3.1) and universal PoS tags (Section 3.3). All feature values were standardized to zero mean and unit variance, as described in Section 3.5.

4.5 Results

In order to evaluate and compare the four baselines described above, they were implemented and tested on the English, Chinese, and Swedish texts from the BC, LCMC, and SUC corpora (see Section 2.1). The three levels of genre granularity (two, four, and nine classes), combined with data in three languages and two classification directions per language pair, allows for 18 combinations of cross-lingual classification tasks. Two metrics were used for evaluation: Prediction accuracy and the average F1-Score over all genre classes. For a given classifier, let TP_i be the number of correctly predicted texts of genre $i \in C$ (True Positives). Similarly, let FP_i the number of texts that were predicted as i, but actually belong to a different genre (False Positives), and let FN_i be the number of texts of genre i that were predicted to belong to a different genre (False Negatives). The precision and recall for genre i are $P_i = \frac{TP_i}{TP_i+FP_i}$ and $R_i = \frac{TP_i}{TP_i+FN_i}$, respectively. The total number of texts in the test set is defined as

$$N = \sum_{i=1}^{C} (TP_i + FP_i) = \sum_{i=1}^{C} (TP_i + FN_i)$$

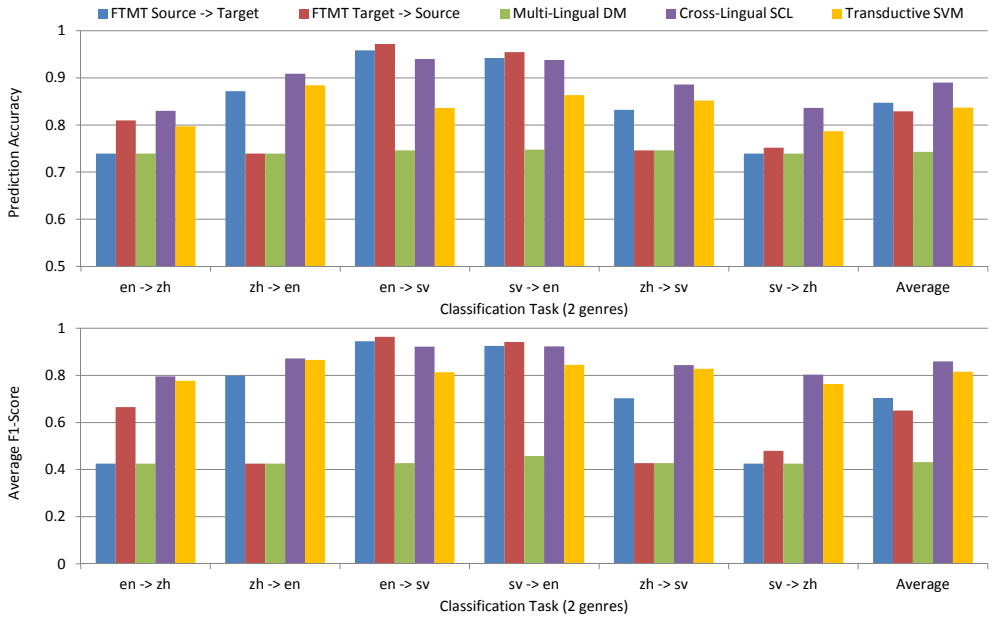

Figure 4.2: Baseline accuracies (top) and average F1-Scores (bottom) for the two genre classification tasks using English, Chinese, and Swedish texts. The graphs show the performances for all six possible cross-lingual tasks, as well as an average.

The classifier's prediction accuracy is then

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{N}$$

and its average F1-Score is

$$F1 = \frac{1}{C} \sum_{i=1}^{C} \frac{2P_iR_i}{P_i + R_i}$$

Note that an F1-Score of 0 was used for a genre class with no predictions. While prediction accuracy is an intuitive and popular metric, which is commonly reported in (text) classification problems, it may be misleading for imbalanced class distributions, such as those shown in Figure 2.1. A classifier can achieve a high accuracy by always predicting the most dominant class, which can be guessed from training data. The average F1-Score is less vulnerable to this, as the recall and precision values of all genre classes affect the score equally, regardless of their number of texts. That is, in order to achieve a high average F1-Score, a classifier must perform well for all classes.

Figures 4.2, 4.3, and 4.4 show the prediction accuracies (top) and F1-Scores (bottom) for the six possible classification directions of the two, four, and nine genre tasks, respectively. Also displayed are the averages over all six tasks for each classifier, metric, and level of granularity. Note that the y-axes differ for all of these graphs. Some

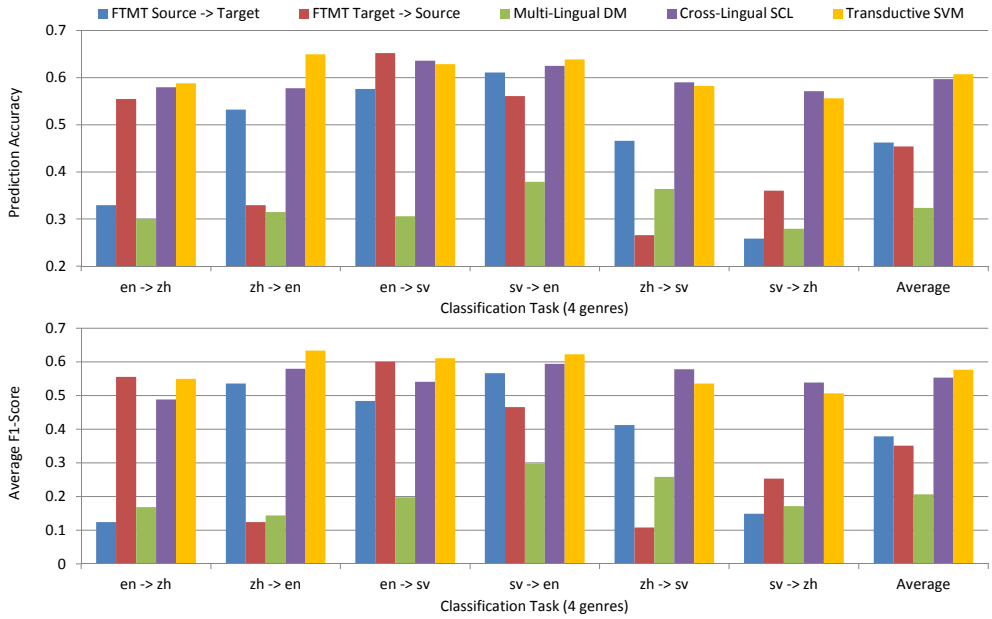

Figure 4.3: Same as Figure 4.2, but for the four genre tasks.

observations can be made from this. Firstly, the multi-lingual domain models proposed by Gliozzo and Strapparava (2006) achieve a very low performance (green bars), as could be expected for the reasons explained in Section 4.2. While for most tasks, the approach outperforms a random guess baseline, this is often due to overprediction of dominant genre classes, which explains the poor average F1-Scores. Unsurprisingly, the method performs a little better in the English-Swedish tasks than the English-Chinese tasks (cf. Figure 4.1). However, for none of the 36 task-metric combinations, it can outperform all other baselines and for most of them, it shows the lowest performance.

The full text machine translation (FTMT; blue and red bars) baseline described in Section 4.1 achieves respectable accuracy in some tasks. It works well for Swedish and English, but worse for language pairs that include Chinese. This is true for both translation directions, but particularly so where translations are made into Chinese. This is likely due to the more difficult word tokenization in these scenarios. The FTMT baseline distinguishes the two broad genres *Informative texts* and *Imaginative texts* very well, in particular for the English-Swedish language pair, where accuracies beyond 95% are achieved. However, accuracy predictably drops for more fine-grained classification tasks. When distinguishing between nine classes for the English-Swedish language pairs, the classifiers average 46.0% across both classification tasks and both translation directions. While this clearly outperforms the accuracy of a random guess prediction

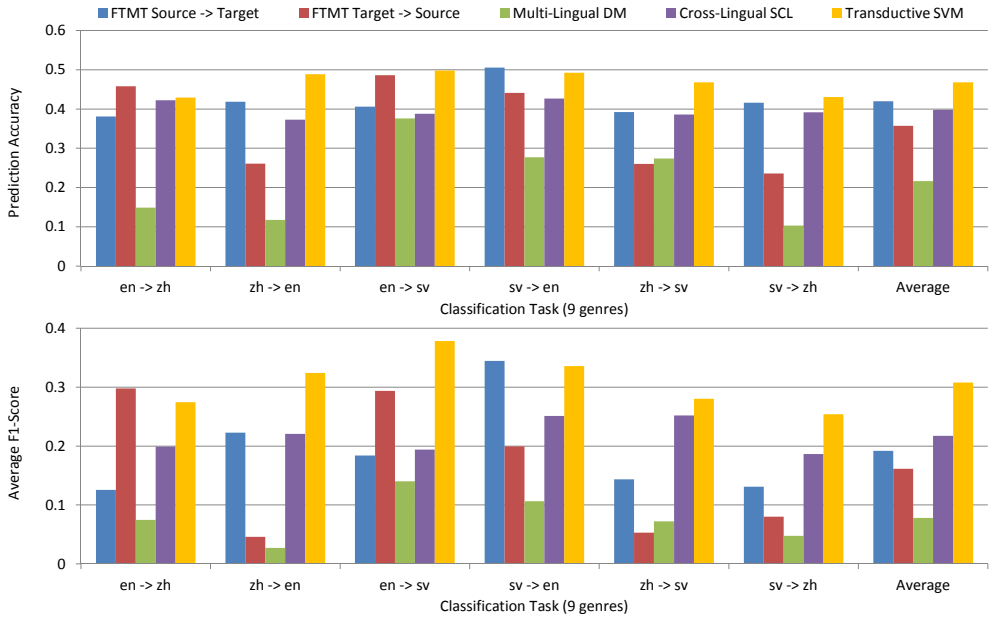

Figure 4.4: Same as Figure 4.2, but for the nine genre tasks.

(11.1%), a detailed look into the results reveals that the classifiers exploit the uneven genre distribution in the data and overpredict dominant classes. For example, not a single text is classified as either biography, editorial, popular lore, reportage, or review for the Swedish to translated English task. Conversely, 295 texts are predicted to be scientfic writing, of which only 77 are true positives. This effect is even stronger in language pairs that include Chinese. For example, the Chinese to translated English classifier simply labels all texts as fiction and outperforms a random guess only due to the strong presence of this genre in the BC (cf. Figure 2.1). This is the reason why the FTMT baseline performs relatively poorly when evaluated by average F1-Scores, in particular for the nine genre task (Figure 4.4).

The cross-lingual SCL algorithm (purple bars) proposed by Prettenhofer and Stein (2010) achieves the best average performance (both accuracy and average F1-Score) of all baselines for the broad two genre task. Here, too, the strongest results can be observed for the English-Swedish classifiers. This difference disappears, however, when genre granularity is increased. For the nine genre task, accuracy is similar across all three language pairs and both classification directions. Furthermore, the cross-lingual SCL classifier achieves a slightly lower accuracy and much lower F1-Score than the TSVM baseline for all six combinations. Compared to FTMT, it achieves more stable, and often better results in spite of a lower use of cross-lingual resources.

The TSVM baseline (orange bars) outperforms all other approaches when averaging over the six medium-grained and fine-grained tasks. This is true for prediction accuracy, but even more for average F1-Score. While it also performs well in the broad-grained two genre tasks, the cross-lingual SCL method achieves better results. While the TSVM method uses less cross-lingual resources than the other baselines, except for the multilingual domain models, it is the only approach to use a genre-specific feature set, rather than word frequencies. The comparatively good results may be explained with the results of Petrenz (2009), who found that word frequencies are poor genre features if the topic-genre correlation is different in the training and in the test set. While that is not necessarily the case here, such correlations are likely to differ in texts from different languages.

Chapter 5

Iterative Re-labelling

As mentioned in Section 1.4, the semi-supervised approaches to CLGC that were developed in this project rely on two strategies: *Cross-Lingual features* and *target language adaptation*. The former is based on the assumption that certain features are indicative of certain genres in both the source language L_S and the target language L_T. The latter is a less restricted way to boost performance, once the language gap has been bridged. An iterative re-labelling algorithm, which is based on these two principles, is presented below. This chapter is based on the work reported by Petrenz (2012), although minor adjustments to the algorithm have since been adopted and more extensive empirical results have been obtained.

5.1 Method Overview

The iterative re-labelling approach is similar to the cross-lingual text classification algorithm proposed by Rigutini et al. (2005). Both methods initially exploit the labelled texts in L_S to predict classes in a set of previously unlabelled target language texts. These predicted labels are then used to iteratively re-train a classifier and update the predictions of the texts in L_T. However, there are differences. Most importantly, while Rigutini et al. (2005) translate all L_S texts into L_T before training the first classification model, no translation is required for the iterative re-labelling method proposed here. Instead, the L_S texts are transformed into a cross-lingual feature representation, as described in Chapter 3. Feature values are then scaled (see Section 3.5) and used to train a supervised classifier. This is used to predict labels for a set of L_T texts, which can be represented in the same cross-lingual feature space. In other words, the language gap is bridged by features which correlate with genres similarly in different languages,

rather than the use of machine translation as in (Rigutini et al. 2005).

Once an initial prediction is made, a new classifier can be trained from L_T texts, based on the newly assigned labels. This is called target language adaptation (TLA) in this project. It is based on the assumption that cross-lingual prediction provides a good but enhanceable result, that is significantly below mono-lingual performance. The resulting decent, though imperfect, genre labels of L_T texts may be exploited to further improve classification accuracy. However, not all texts are assigned to a genre with equal confidence, as some texts are easier to classify than others. Ideally, the new classifier should be trained only from texts that have high confidence in their labels, especially in early iterations, where uncertainty may be high. For some classifiers, such as Naïve Bayes, this confidence can be directly derived from the posterior class probabilities. That is, the higher the posterior probability for the predicted genre class, the higher the confidence of the classifier in its prediction for a given text. Other classification methods, such as SVMs, do not compute posterior probabilities, but classify texts based on which side of a previously learned decision hyperplane they fall. In that case, the distance of a text to this hyperplane can be used to infer a prediction confidence value. Once a suitable subset of texts has been chosen, a new classification model is trained and used to update the labels of all L_T texts. This process is repeated until convergence, that is until the predicted genre labels for all L_T texts are identical in two subsequent iterations.

One advantage of making the separation between the cross-lingual prediction and the subsequent TLA is that a different set of features can be used for the latter. There is no reason to keep the restrictions required for cross-lingual features once the language gap has been bridged, as L_S texts are not involved in the following steps. In fact, any feature that is known to correlate with genre in L_T may be used for this task. However, since the aim of CLGC is to bring the benefits of genre classification to a language without appropriate training data, it is likely that there is little or no knowledge about such L_T specific features. Therefore, it makes sense to use automatically generated feature sets such as word frequencies, which do not require knowledge of L_T. This set is called *TLA features* in this chapter.

Rigutini et al. (2005) also showed that supervised feature selection can help to prevent trivial solutions and thus increase classification accuracy. Here, their suggestion of ranking features by their information gain is adopted. In each iteration of the TLA process, the top k features in this ranking are selected and used for training the new classification model. Note that the information gain is computed based on the chosen

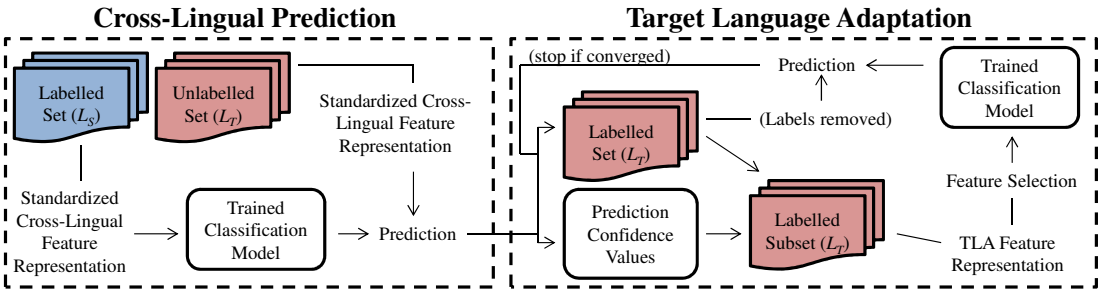

Figure 5.1: Outline of the iterative re-labelling algorithm with its two components: Cross-Lingual Prediction and Target Language Adaptation. Source language texts are marked blue, target language texts are marked red.

high-confidence texts only and that its computation is repeated in every iteration. Note also that, unlike in (Rigutini et al. 2005), the features for the initial cross-lingual prediction are not selected in this fashion. This is because information gain can only evaluate the predictive power for features within L_S. It cannot estimate how well a feature is suited for a cross-lingual task for the lack of labels in L_T.

Figure 5.1 illustrates the complete algorithm, which can be described in eight steps. To this end, let S be the set of labelled L_S texts and let T be the set of unlabelled L_T texts.

1. Extract cross-lingual feature values from the texts in S and standardize.
2. Extract cross-lingual feature values from the texts in T and standardize.
3. Train a classification model on S to predict genre labels for all texts in T.
4. Select a subset $T' \in T$ with high prediction confidence (see Section 5.2).
5. Based on T', compute an information gain ranking on the full set of TLA features.
6. Represent all texts in T as a set of the top k features from this ranking.
7. Train a classification model on T' to predict genre labels for all texts in T.
8. If the predicted labels are identical to those of the last iteration, stop. Otherwise, go to step 4.

Iterative re-labelling is a wrapper algorithm that can be applied to a number of different machine learning methods, as long they provide a way to induce prediction confidence. After initial experiments with Naïve Bayes, k-nearest-neighbours, and Support Vector Machines, the latter were found to perform best in this setting. Therefore, this chapter focusses on SVM models as the classifier and results are reported accordingly.

5.2 Prediction Confidence

The decision hyperplane of a trained SVM model in a binary classification problem is defined by the weight vector $\mathbf{w}$ and the intercept term b. From this, the Euclidean distance of a point $\mathbf{x}$ to the hyperplane can be computed as

$$d = \mathbf{w}^T \cdot \mathbf{x} + b$$

The larger $|d|$ is, the more confident the classifier is in its prediction for $\mathbf{x}$. Consider the example illustrated in Figure 5.2. It shows a scatterplot of the 483 English texts in the BC based on their standardized average word lengths and noun frequencies, with black marks denoting informative texts and red marks denoting imaginative texts. It also shows the decision boundary learned by a linear SVM classifier that was trained on the Chinese texts from the LCMC and the same two features. There is a fairly high amount of confusion close to the decision boundary, that is black marks on green background and red marks on yellow background, both of which stand for misclassifications. However, for texts that are far from the boundary, little or no confusion can be observed. In other words, the further a text is from the decision hyperplane, the lower the probability that it was classified incorrectly. The distance d can therefore be used to rank L_T texts in descending order and select only a highly ranked subset for training the next iteration's classifier.

In the experiments for this project, the SVM implementation of the LIBSVM library (Chang and Lin 2011) was used. For multi-class tasks, this library employs a one vs. one approach, that is for a problem with $|C|$ genre classes, $|C|(|C|-1)/2$ binary classifiers are trained, which are then used to vote on the prediction of a new text. This of course means that there are $|C|(|C|-1)/2$ decision boundaries and $|C|(|C|-1)/2$ distances for each L_T text. Therefore, the following strategy was adopted. First, only the $|C|-1$ binary classifiers where one class is the predicted genre (based on the overall voting) were considered. The distances for a given L_T text to the $|C|-1$ decision hyperplanes are stored in a vector $\mathbf{d} = \{d_1, ..., d_{|C|-1}\}$. One way to get a prediction confidence value for this text would be to take the sum of this set, that is

$$\sum_{i=1}^{|C|-1} d_i$$

Since the result will be used to rank texts, this is equivalent to the arithmetic mean. However, the product of the set seemed a better choice, that is

$$\prod_{i=1}^{|C|-1} d_i$$

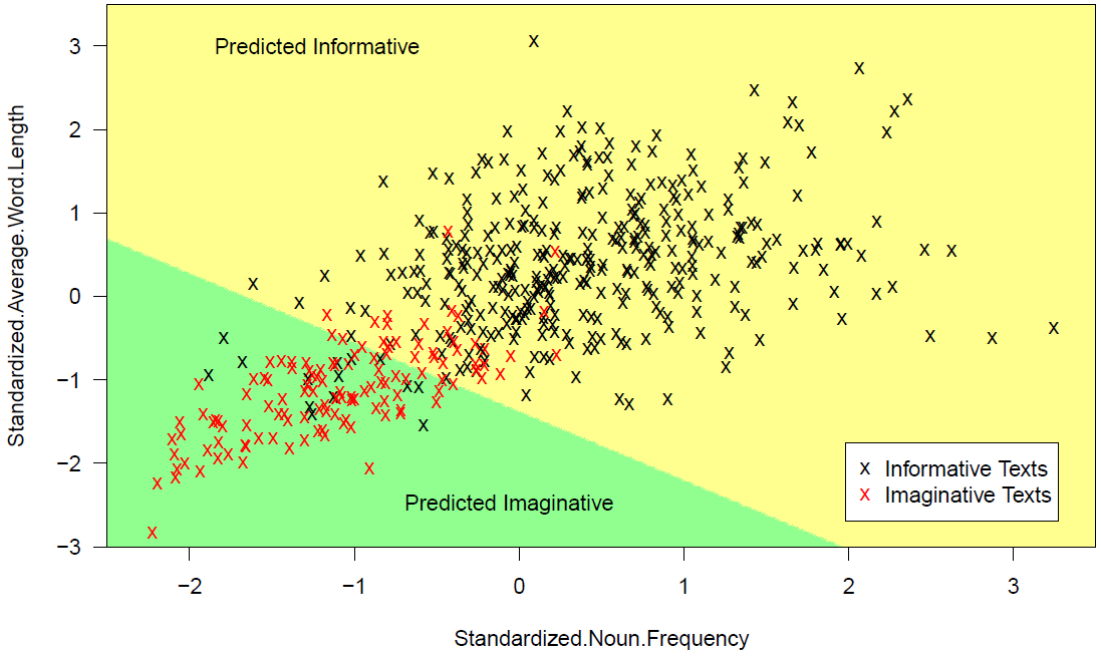

Figure 5.2: Visualization of a linear SVM classification model trained on Chinese texts using average word lengths and noun frequencies as features. The marks denote English texts (black: informative; red: imaginative). Texts that fall into the yellow and green areas would be classified as informative and imaginative texts, respectively.

This is equivalent to the geometric mean. The reason is that the geometric mean, unlike the arithmetic mean, heavily punishes very small values, that is very small distances to the decision hyperplane. The intuition is that a confidently classified text ideally should be far away from the decision boundary for all one vs. one classifiers.

However, for a given binary classifier i the distance d_i may be negative, that is the classifier's prediction differs from that of the overall voting. In order to avoid negative confidence values, or rewarding high negative values, these distances are transformed to a small positive value. This should be small enough to reduce the overall confidence value, since disagreement of different classifiers is undesirable. It should, however, not be zero, as the confidence should still be comparable and rankable, based on the distances to other decision hyperplanes. In the experiments for this project, 10^{-5} is used as a value for this parameter.

Another decision is the number of texts selected to train the next classifier. Generally speaking, this is a trade-off between a training set which has little confusion (small subset) and a sufficiently large number of training texts (large subset). Since some genres are easier to identify than others, the confidence values of their texts may be

comparatively high. Thus, simply choosing the overall top ranked texts may lead to a very imbalanced genre distribution in the new training set. Rather, in this project, a percentage of texts in each previously predicted genre class was selected. That is, for each predicted class, the top $p\%$ of confidence ranked texts were used to train the next classifier. Furthermore, the choice of p can be different in each iteration. Intuitively, uncertainty should be high in the early iterations, in particular for the initial cross-lingual prediction, as the classifier can only exploit a limited set of features and distributions may be different in L_S and L_T. It therefore makes sense to use a relatively low value for p in the first iteration and increasingly higher values later on. Accordingly, for the experiments of this project, p was chosen to be 0.6 after the cross-lingual prediction (p_1), and 0.9 for the second iteration (p_2). From there, it was increased by 0.05 in each iteration, up to a maximum of 1.0. Note that these values were chosen based on intuition, rather than supervised feature optimization, as the latter would require genre annotated texts in L_T. However, Section 5.3 includes an evaluation of different values for p_1 and p_2.

The choice for the value of k (the number of features used in each iteration of the TLA process) can either be determined statically or dynamically by using an information gain threshold. In the experiments of this project, a static value of 500 dimensions was used.

Note that confidence values can be obtained from SVMs in different ways, such as the regression method by Platt (1999), who uses a sigmoid function to map SVM outputs into probabilities. This would not make a difference in ranking the texts by confidence, and therefore would not affect the results of the method as described in this section. However, if the confidence values were instead used to determine label weights to inform the training process of the next iteration's classifier, Platt's method might be a sensible choice. This was not employed in this project, but could be an interesting extension for future work.

5.3 Results

To assess the performance of the iterative re-labelling algorithm, experiments were carried out with different source and target languages. Unless otherwise stated, the method exploits only text statistics for the initial cross-lingual prediction and L_T word frequencies during the TLA process. In particular, note that (unlike in Chapter 4) no universal PoS tag features were used for the results presented here, with the exception

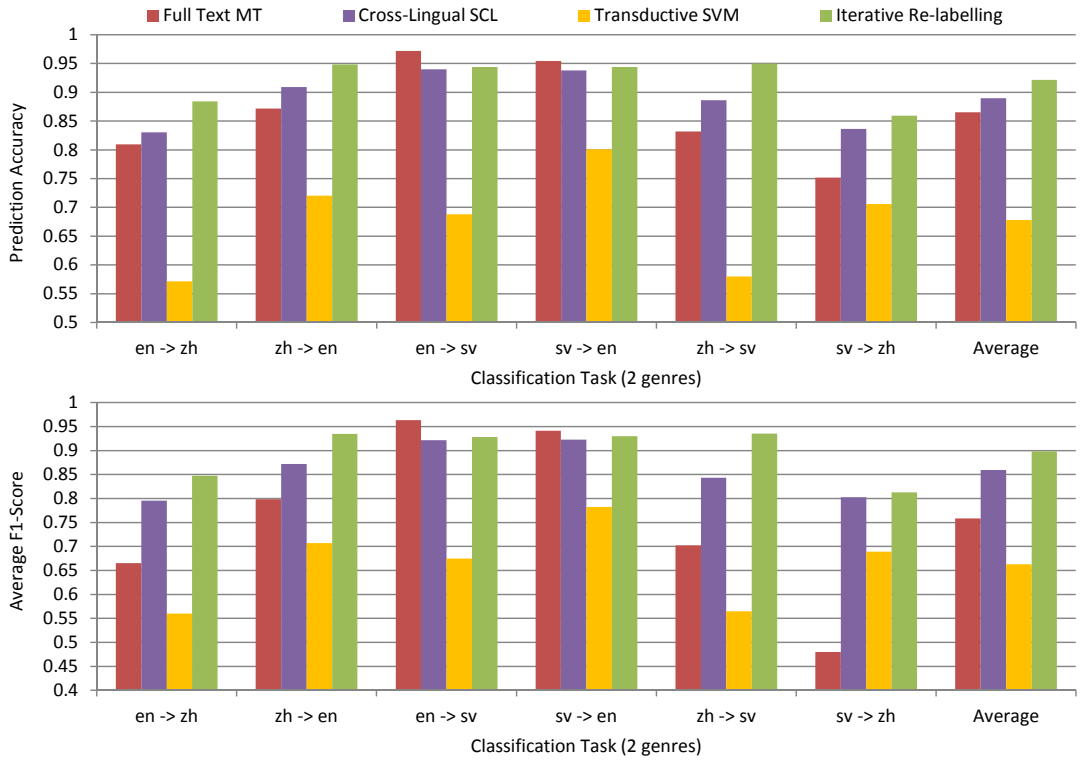

Figure 5.3: Prediction accuracies (top) and average F1-Scores (bottom) for the six classification tasks with two target genres using English (BC), Chinese (LCMC), and Swedish (SUC) texts. The full text MT baseline bars (red) illustrate the better result of the two possible translation directions for each task. The transductive SVM baseline (yellow bars) and the iterative re-labelling method (green bars) use the same cross-lingual feature set (text statistics).

of Figure 5.8. The approach was evaluated on the data from the BC (English), LCMC (Chinese), and SUC (Swedish) corpora described in Section 2.1. Figures 5.3, 5.4, and 5.5 show the classification results for the tasks with two, four, and nine genres, respectively. The baselines shown are the ones described in Chapter 4, except for the multi-lingual domain models, which performed poorly on all tasks. Note that the full text machine translation results shown are somewhat artificial, as the better result of the two possible translation directions is displayed for each task. For easier comparison, the TSVM baseline here uses the same cross-lingual feature set (i.e. text statistics) as the iterative re-labelling method and therefore results differ from those displayed in Figures 4.2, 4.3, and 4.4. An evaluation of how much the different methods can benefit from additional features will be presented later in this Section.

For the two and four genre tasks, the iterative re-labelling method achieves very

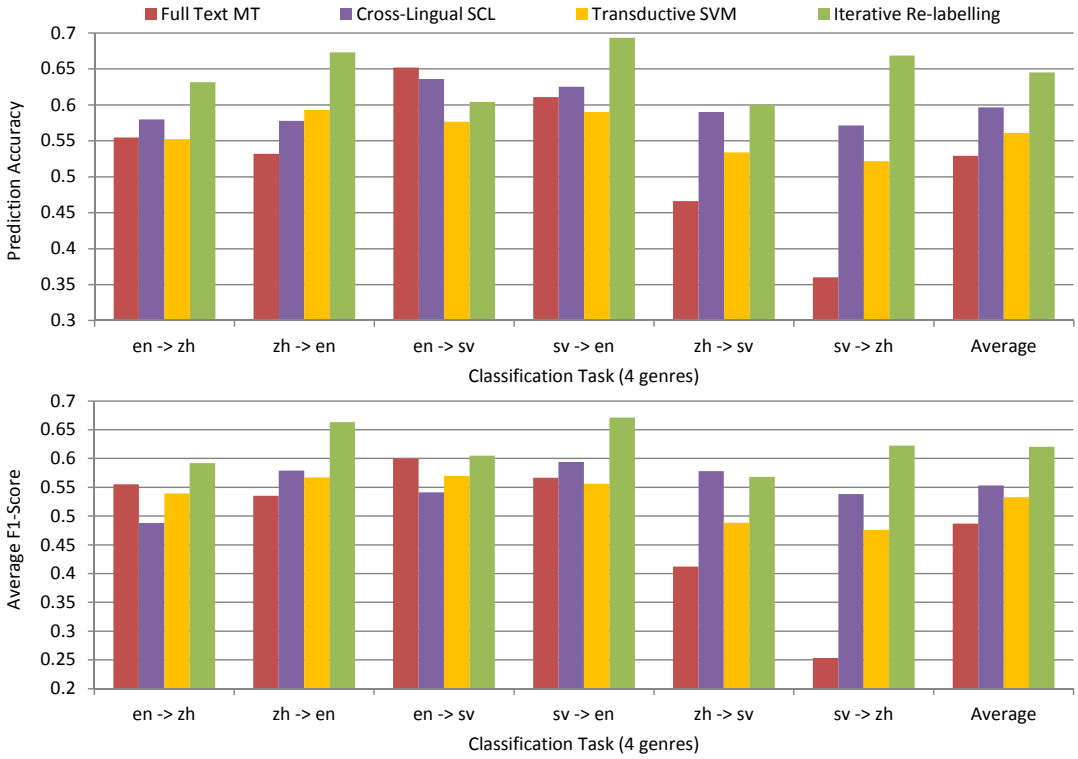

Figure 5.4: Same as Figure 5.3, but for the four genre tasks.

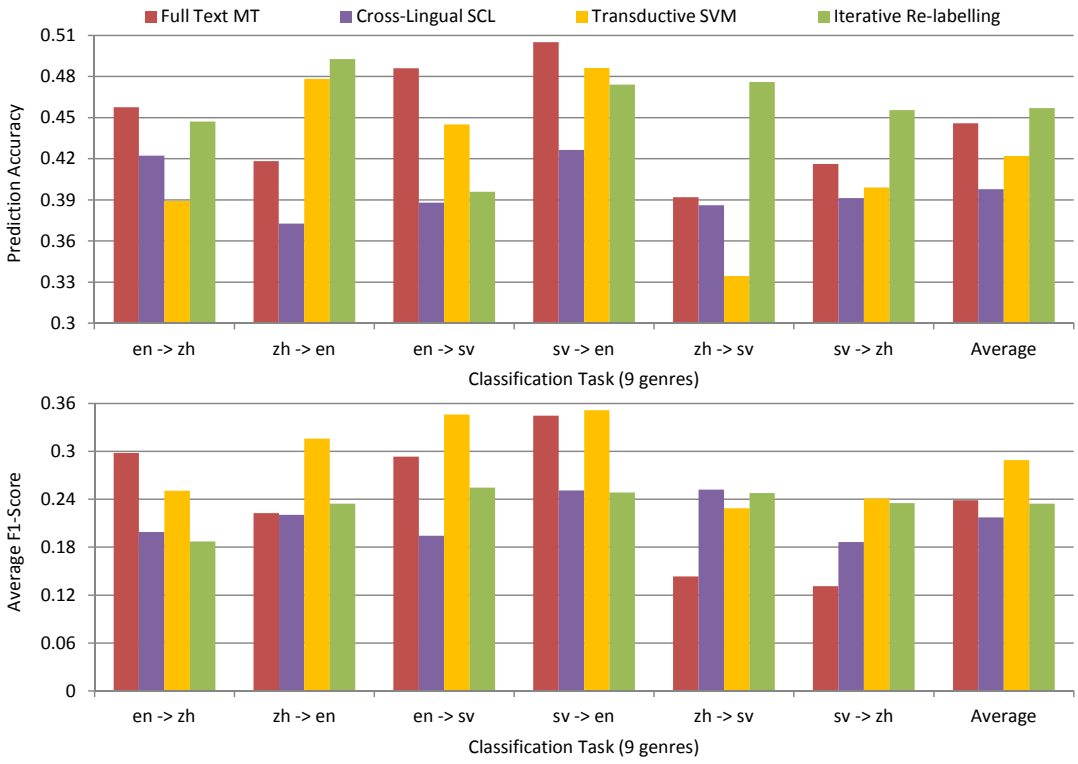

Figure 5.5: Same as Figure 5.3, but for the nine genre tasks.

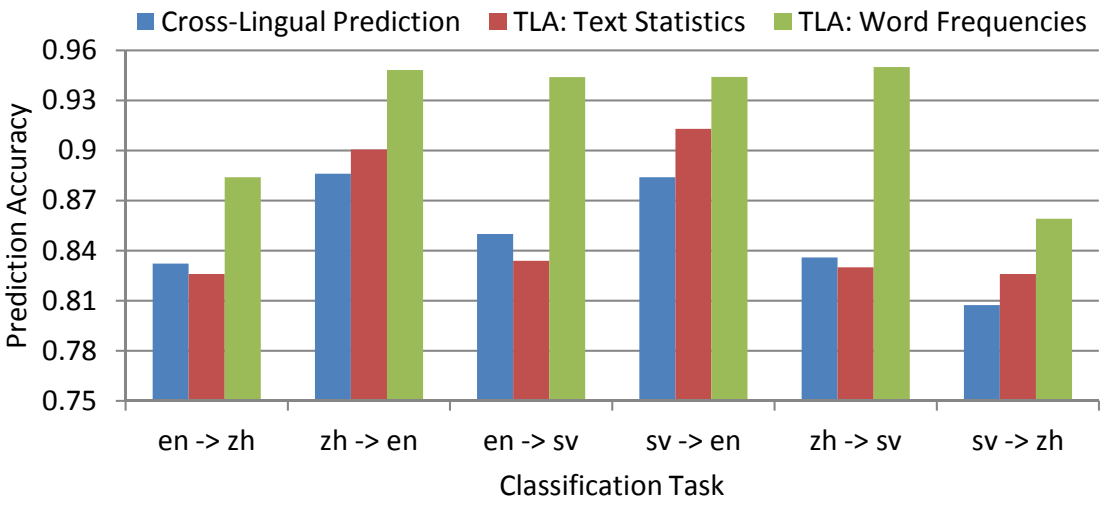

Figure 5.6: Prediction accuracies for the six binary classification tasks. Blue bars show the accuracy after the cross-lingual prediction. Red and green bars show the accuracy after TLA convergence, exploiting L_T word frequencies and text statistics, respectively.

competitive results. In both the accuracy and the F1-Score metrics, it outperforms all baselines for four and five tasks in the binary and four-class problems respectively. Averaged over all six tasks, it achieves the highest accuracy and F1-Score for both levels of genre granularity. This is remarkable, since the method uses no cross-lingual resources, unlike the FTMT and CL-SCL baselines, which rely on the availability of a machine translation system. The TSVM baseline requires no such resources either, but performs significantly worse in the two and four genre tasks.

In the nine genre task, the performance of the iterative re-labelling method is less outstanding, albeit still impressive. For three of six tasks it achieves the highest prediction accuracy, but is outperformed by the FTMT baseline where English and Swedish texts are used. It also does not achieve the best average F1-Score for any of the six tasks, but overall is comparable to the FTMT and CL-SCL baselines on this metric. The TSVM baseline achieves the highest F1-Scores for most tasks.

These results demonstrate that good results in a CLGC problem are possible to achieve without extensive linguistic resources and/or machine translation. In fact, the experimental outcomes show that for most tasks, the iterative re-labelling method performs equally well or better than more resource-hungry techniques. This is particularly true where coarser genres are used as target classes but less obvious for the more fine-grained problem. This is likely due to the relatively small cross-lingual feature set used.

As mentioned in Section 5.1, one benefit of the iterative re-labelling method is that different features can be used during the cross-lingual prediction and TLA phases. Figure 5.6 illustrates this benefit. It shows the prediction accuracies for the two genre tasks after the initial cross-lingual prediction. It also shows the performance for two different feature sets used in the TLA process – text statistics and word frequencies. The former is the same set of features used in cross-lingual prediction, while the latter can only be used within L_T, due to the differences to the L_S vocabulary.

It is clear that the classifier benefits from this larger, less restricted set of TLA features. For all six tasks, the word frequency features outperform the text statistics. In fact, accuracy does not generally improve during the TLA process if the same features as in cross-lingual prediction are used. On the other hand, with L_T specific features, the algorithm manages to exploit the structure of the unlabelled L_T texts and improve results compared to the initial accuracies for all six L_S-L_T combinations. This shows that it can be beneficial to include genre-revealing features for L_T in a semi-supervised classification method, even if they cannot be used across languages. It is also further evidence that what works well across languages is not necessarily the best choice within L_T, and vice versa.

The iterative re-labelling method uses the distance of texts from the decision hyperplane in order to select a subset for training the next classification model. Figure 5.7 shows an evaluation of different parameter settings, which determine the size of this subset in different iterations of the algorithm. The value for p_1 corresponds to the percentage of texts in each cross-lingually predicted genre class, therefore it determines the size of the first TLA training set. The value of p_2 is the equivalent for the second TLA training set, that is after the first prediction of a L_T-trained classifier. The value of p increases by 0.05 in each subsequent iteration. For the evaluation, 10 values from 0.1 to 1 for p_1 where tested with p_2 set to 1, and vice versa.

The upper left graph of Figure 5.7 demonstrates that the classifier strongly benefits from omitting the low-confidence L_T texts from the training set after the cross-lingual prediction step. For all six L_S-L_T combinations, all tested values $0.1 < p_1 < 1$ achieved a higher accuracy than $p_1 = 1$, which corresponds to no confidence-based selection. In fact, performance starts dropping for $p_1 > 0.7$, or even before in some cases. The upper right graph shows a similar trend for p_2, as the results of all tested $0.6 < p_2 < 1$ are better than that of $p_2 = 1$ for all six tasks. However, lower values of p_2 seem to harm classifier performance, in particular where Chinese is the target language. The bottom graph illustrates this difference between p_1 and p_2, as the average values of the upper

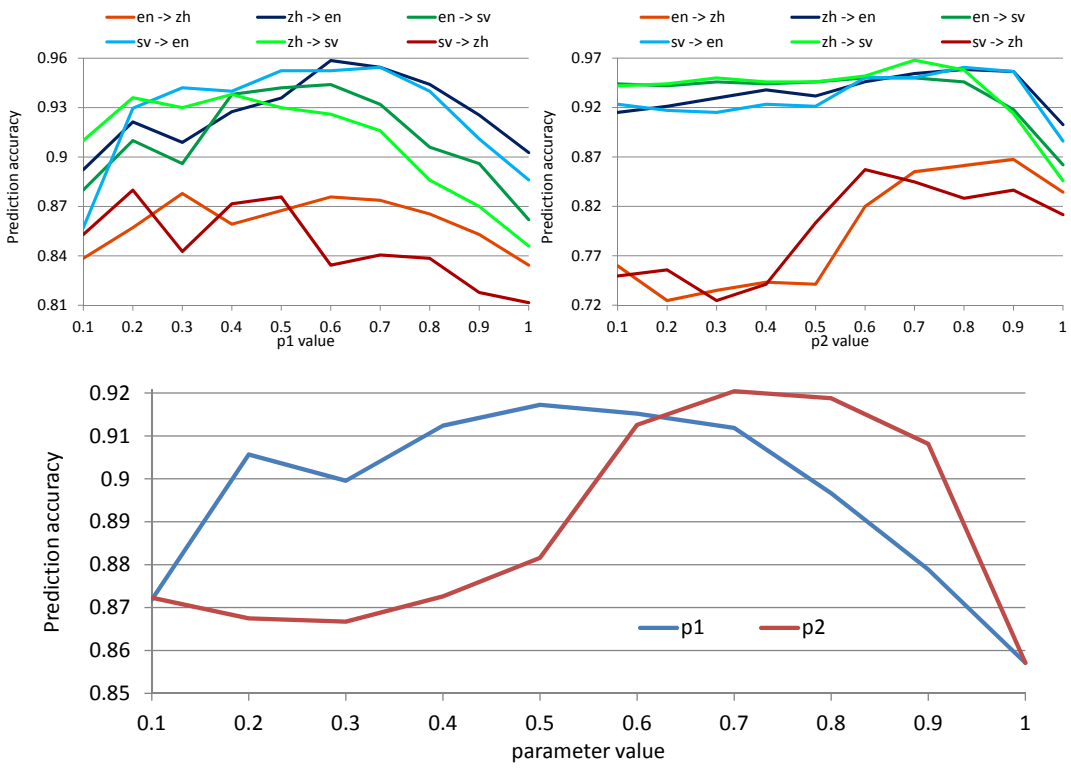

Figure 5.7: Prediction accuracies for the binary classification tasks for different parameter values. Upper left: Parameter p_1 ranges from 0.1 to 1, parameter p_2 set to 1. Upper Right: Vice versa. Lower: Averages of the values in the two upper graphs (blue: upper left; red: upper right) over all six classification tasks.

two graphs are shown. This demonstrates the benefit of a relatively strict selection of texts after the cross-lingual prediction, where confusion is predictably high. Once a L_T based classification model was used to predict genres, a more generous selection is appropriate in order to keep training sets sufficiently large. Even so, omitting a small percentage of texts with low prediction confidence in the early iterations of the TLA process can improve accuracy. As can be seen, the intuitively chosen values of $p_1 = 0.6$ and $p_2 = 0.9$ are reasonable, although not optimal judging by this evaluation.

The algorithm as evaluated above uses a minimum of knowledge about L_T. Since often further knowledge and resources are available, it would be desirable if this could easily be integrated into a classifier in order to improve results. If, for example, L_T data exists that allows training a supervised PoS tagger, additional cross-lingual features can be exploited by mapping tags to a universal set, as explained in Section 3.3. Figure 5.8 shows an evaluation of whether the iterative re-labelling method or the TSVM baseline can benefit from this additional resource. The light blue and light red bars are the same

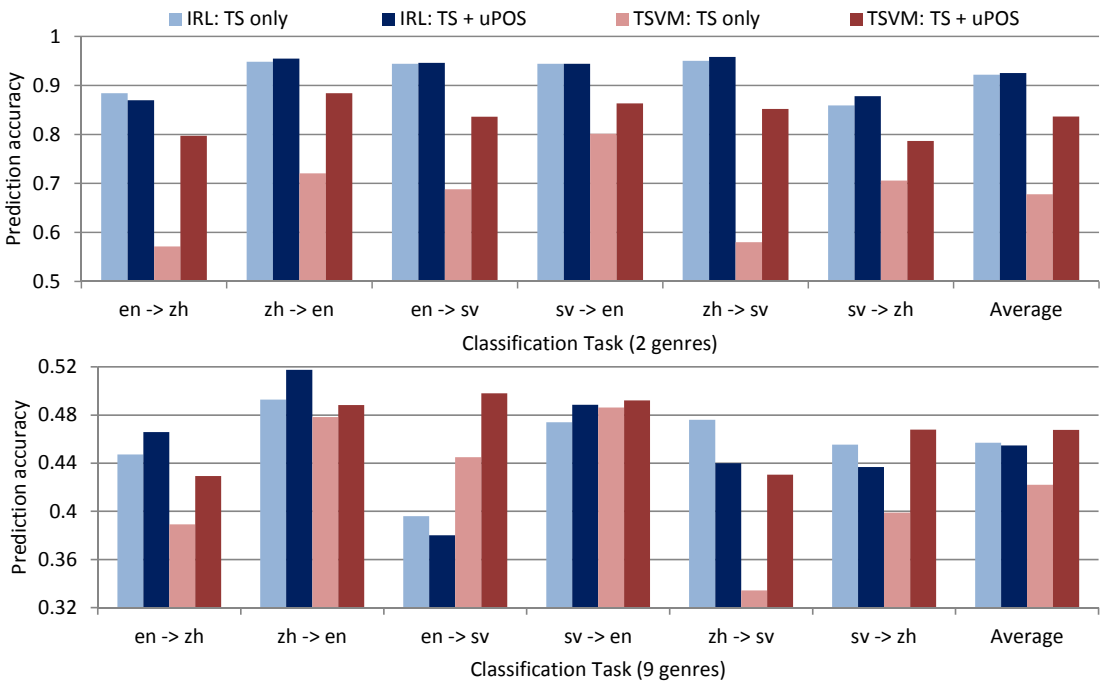

Figure 5.8: Prediction accuracies for the two (upper) and nine (lower) genre classification tasks. The results are achieved by the iterative re-labelling (IRL) method using only text statistics (light blue) and a combination of text statistics and universal PoS features (dark blue) in the cross-lingual prediction phase. The light and dark red bars denote TSVM baseline results for the same feature sets, respectively.

results already shown in Figures 5.3 (2 genres) and 5.5 (nine genres). The dark blue and dark red bars are the result if both text statistics and universal PoS frequencies are used in the cross-lingual set. Note that the iterative re-labelling method still exploits word frequencies only during the TLA process. For the TSVM baseline, these results are the same already reported in Figures 4.2 and 4.4.

It is clear that the TSVM baseline benefits strongly from the additional PoS-based features. For all L_S-L_T combinations and for both binary and multi-class problems, accuracy improved due to these new predictors. The iterative re-labelling method, on the other hand, achieves similar results with the universal PoS frequencies as it did without them. While for some tasks, slightly improved accuracies were observed, the opposite is the case for other tasks. The average accuracies over all tasks are nearly identical. It might be argued that for the two genre task (upper graph in Figure 5.8), the difference between the two algorithms is due to the much lower starting point of the TSVM classifier. That is, one reason for the lack of improvement of the iterative re-

labelling method with the additional features might be due to the already high accuracy. However, the algorithm, unlike TSVM, does not benefit from the new predictors even for tasks with low accuracy, such as the nine genre en → sv problem.

This provides evidence that the cross-lingual SVM model used to initially predict genres for L_T texts may not benefit from additional linguistic resources, should they be available. Since it establishes decision hyperplanes based on L_S texts only, it has no way of weighting features based on their predictive power across languages. It relies on the correlations between feature values and genres to be similar in L_S and L_T. While the results presented here show that this is a reasonable assumption for some of the features, a more robust classifier would ideally be able to select or weight predictors based on both L_S and L_T, rather than just L_S. This way, additional features may be exploited to improve results, assuming they have predictive power with regards to the target genres.

In conclusion, the method introduced in this chapter achieves strong results by exploiting (1) a small set of text statistic features to bridge the language gap, and (2) a set of L_T word frequencies to iteratively re-label target texts. Both prediction accuracy and average F1-Score are particularly impressive for smaller sets of broad target genres and equal or better than those of more resource-intensive baselines, thus providing evidence for the thesis described in Section 1.4. The method owes is strong results partly to the different sets of features it employs in different stages of the algorithm. It also benefits from the computation of prediction confidence values and the selection of texts for training classification models. However, the lack of cross-lingual feature weighting may diminish the method's ability to benefit from additional cross-lingual resources or linguistic tools in L_T.

Chapter 6

Exploiting Comparable Corpora

The results presented in Chapter 5 suggest that simply extractable features can be predictive of genre across languages. However, the employed supervised cross-lingual classification model has no way of identifying which features are particularly helpful for bridging the language gap. It can only find a classification boundary based on training data in the source language, and hope that feature values correlate with genre similarly in the target language. This seems to work well for the proposed text statistics feature set. Nevertheless, a classification model that can automatically identify features which work well across languages may further improve performance, especially as further resources become available.

Intuitively, this requires labelled data in the target language which is assumed to be unavailable in this project. Genre-annotated texts may however be available in more than one source language. This can potentially be exploited in order to construct more robust cross-lingual classifiers. Two such approaches are presented, evaluated, and discussed in the following sections. They are based on the work in (Petrenz and Webber 2012b).

6.1 Method Overview

The methods in this chapter exploit comparable corpora to improve classifier performance. This denotes a collection of similar texts in different languages, which are not translations of each other. Typically, this similarity is defined by subject matter, that is texts are comparable in their topic. For the purpose of this project however, similarity is defined by genre. In other words, the methods presented here require genre-annotated texts from a comparable set of genres in different languages, while the subject matter

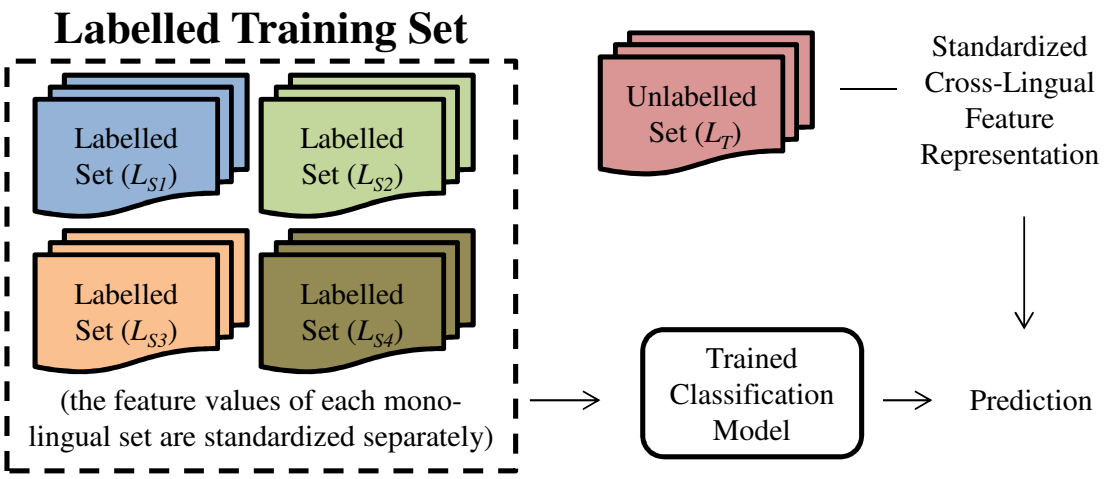

Figure 6.1: Outline of the multi-lingual training approach to exploiting a comparable corpus for CLGC. In this example, a set comprising texts in four different languages is used to train the cross-lingual classifier.

may differ. Note that, unlike the method in Gliozzo and Strapparava (2006), discussed in Section 4.2, the approaches proposed here do not require the comparable corpus to include texts from the target language.

Comparable corpora have the potential to benefit CLGC in two ways: First, they allow sets of text written in different languages to be combined into a single set that can be used to train a classification model. The hypothesis is that such a multi-lingual training set should make the model more robust for cross-lingual classification tasks than a mono-lingual training set whose genre-related differences might not be the same as those in the target language. This is because any learned classification model will give high weight to features only if their values correlate with genres in general, rather than in one language. Adding more languages to the training set will therefore result in a classification boundary which can separate genres in multiple languages. This makes it more likely to perform well on the target language. Note that no target language texts are assumed to be part of the training set.

The approach is outlined in Figure 6.1. It is an extension to the cross-lingual prediction part of the iterative re-labelling method described in Chapter 5. Since the subsequent TLA part relies on a good initial prediction in order to improve results, any improvement here can be assumed to have further benefits during the iterative process.

Secondly, comparable corpora might allow the automatic identification of features, which perform well across languages. Using a supervised feature selection technique

on a set of several languages may yield features that have predictive power in more than one language. The hypothesis is that using such features can prevent a classifier from overfitting to the idiosyncrasies of the training language, even if that language is not used in the feature selection process. As noted in Section 1.3, cross-lingual text classification can be regarded a special case of a domain adaptation problem, where feature selection techniques have been applied successfully before (Pan et al. 2010). Here, a simple feature-ranking method is employed that uses information gain to determine the value of a feature to predict genres. Information gain is defined as

$$IG(Class, Feature) = H(Class) - H(Class|Feature)$$

where $H(X)$ is the entropy of variable X. In this project, the information.gain function of the R library FSelector (Romanski 2013) was used. This internally discretizes the continuous feature values to obtain nominal variables, using the multi-interval method proposed by (Fayyad and Irani 1993), in order to estimate entropies.

The information gain is computed on a set of texts written in several languages, taken from the comparable corpus. A subset of features can then be obtained by choosing the top k features in this ranking of n features. While the availability of domain knowledge would allow this parameter to be set manually, it can be determined automatically. To this end, the maximum cross-validation accuracy on the comparable corpus is found, where each fold corresponds to training on a single language and testing on all remaining languages. This involves an exhaustive search over all possible values of k. However, using the information gain ranking greatly reduces the possible numbers of feature subsets from $2^n - 1$ to n.

Figure 6.2 illustrates this method. Note that neither the source nor the target languages are represented in the set used for feature ranking and threshold determination.

6.2 Experiments

Evaluating the potential benefits of multi-lingual training sets and cross-lingual feature selection requires a comparable corpus with texts from as many languages as possible. Therefore, the experimental framework used the Reuters (newswire text), Europarl (transcribed speech), and JRC-ACQUIS (legal text) corpora, which share eight European languages (for more details, see Section 2.3).

The cross-lingual features used were text statistics (Section 3.1) and punctuation frequencies (Section 3.2). To these, the frequencies of the 25 most common words in

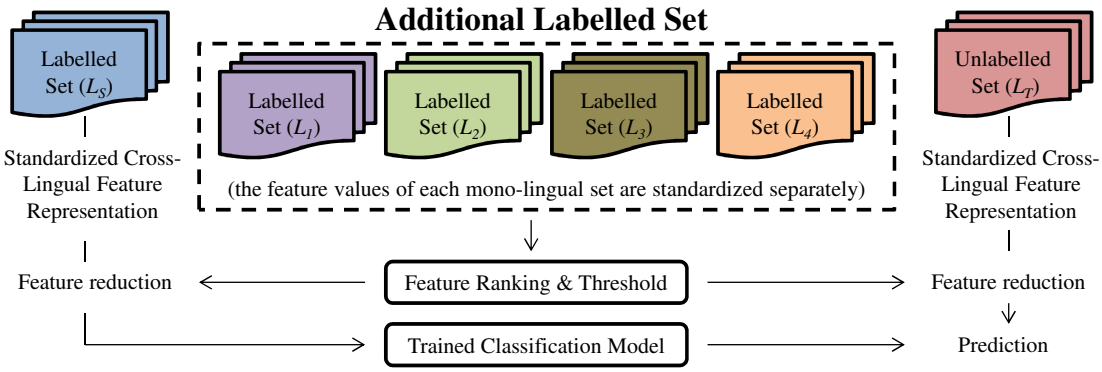

Figure 6.2: Outline of the cross-lingual feature selection approach to exploiting a comparable corpus for CLGC. In this example, a set comprising texts in four different languages is used to rank and select features. A different source language and the reduced feature set are then used to train the classifier, which predicts the genres of texts in the target language.

the respective language were added. Common word frequencies have been shown to have discriminative power in mono-lingual genre classification tasks (Stamatatos et al. 2000a). However, since the i^{th} most common word in language A differs semantically from the i^{th} most common word in language B, it was expected that these features are of little value for a cross-lingual task and that they might have a negative impact on prediction accuracies. They were added to the set in order to evaluate whether the feature selection method would filter out such predictors. The final set comprised 78 features, three of which were discarded, as they had zero values for all texts in one or more languages.

The experiments were designed to answer the question whether a comparable corpus with genre-annotated texts can be exploited to improve the results of a cross-lingual prediction as described in Chapter 5. Therefore, cross-lingual classifiers trained on a single language were used as baselines. To this end, a separate SVM model was trained for each of the eight mono-lingual sets, using all 75 features. Each model was then tested on the seven languages that were not used to train it. This performance is achievable without the use of a comparable corpus.

6.2.1 Multi-Lingual Training

To exploit the genre labels in more than just one language, the representations of seven language sets were merged into a single training set, with one language held back

for testing. Naturally, the merged multi-lingual training set contained seven times as many texts as any mono-lingual baseline. Since supervised classification results tend to improve with larger training set sizes, this bias was removed by splitting the merged set into seven disjoint training sets, keeping the language and genre distributions intact. Thus, for each target language, the SVM model was trained seven times and evaluated by computing the average accuracy.

Table 6.1 contains the prediction accuracies for the 56 single language training experiments (i.e. baseline performances), as well as the accuracies yielded by the combined multi-lingual training sets. Figure 6.3 shows the same results graphically. For each of the eight target languages, the red bar corresponds to the experimental framework illustrated in Figure 6.1. The average accuracy of the red bars over all target languages is 97.5% compared to 95.0% for that of the black and grey bars. Furthermore, for all eight languages, accuracy based on the multi-lingual training set exceeded accuracy based on any of the seven mono-lingual baselines. This significant (sign test; $p < 0.01$) improvement indicates that the knowledge represented by genre labels in different languages can be exploited to build more robust and high-performing CLGC models. The experimental results show this to be true at least for a sufficiently large number of related languages in the training set.

6.2.2 Cross-Lingual Feature Selection

A second experiment was conducted to evaluate whether a comparable corpus can be used to identify good cross-lingual predictors from a set of candidates. Here, features were ranked by their information gain within a set of texts from six languages. Then, 6-fold cross-validation was used to determine the threshold parameter k. This was done by trying $k = 1$ to $k = n = 75$ and choosing the value for k with the highest cross-lingual cross-validation accuracy. The feature sets of the seventh and eighth languages were then reduced to the resulting subset, and used for training and testing respectively.

Table 6.2 contains the gains and losses in prediction accuracy when using only the top k features, as compared to the full feature set. Figure 6.4 shows the same results graphically. For the 56 tasks, k ranged from 13 to 23, with the majority between 13 and 15. Most classification models benefited from this feature selection step and the average accuracy over all tasks improved from 95.0% to 96.6%. Although in some cases worse results were observed, overall performance based on the reduced feature set was significantly better ($p < 10^{-8}$), according to the sign test.

	da	de	en	es	fr	it	pt	sv	μ
Danish (da)	—	.959	.951	.961	.930	.965	.937	.971	.953
German (de)	.943	—	.925	.934	.897	.957	.933	.954	.935
English (en)	.948	.942	—	.961	.934	.962	.942	.972	.952
Spanish (es)	.960	.920	.952	—	.946	.963	.927	.973	.949
French (fr)	.961	.952	.965	.974	—	.973	.940	.967	.962
Italian (it)	.959	.963	.955	.962	.948	—	.949	.953	.956
Portuguese (pt)	.955	.948	.945	.954	.928	.954	—	.961	.949
Swedish (sv)	.965	.949	.948	.963	.911	.947	.928	—	.944
Multi-lingual	**.979**	**.968**	**.973**	**.979**	**.967**	**.980**	**.971**	**.986**	**.975**

Table 6.1: Prediction accuracies for the cross-lingual genre classification tasks. Rows 2-9 denote the training language, Columns 2-9 denote the testing language. The accuracies in row 10 were achieved by training the model on the seven languages which it was not tested on. Column 10 contains the average of each row. The best accuracy for each column is highlighted.

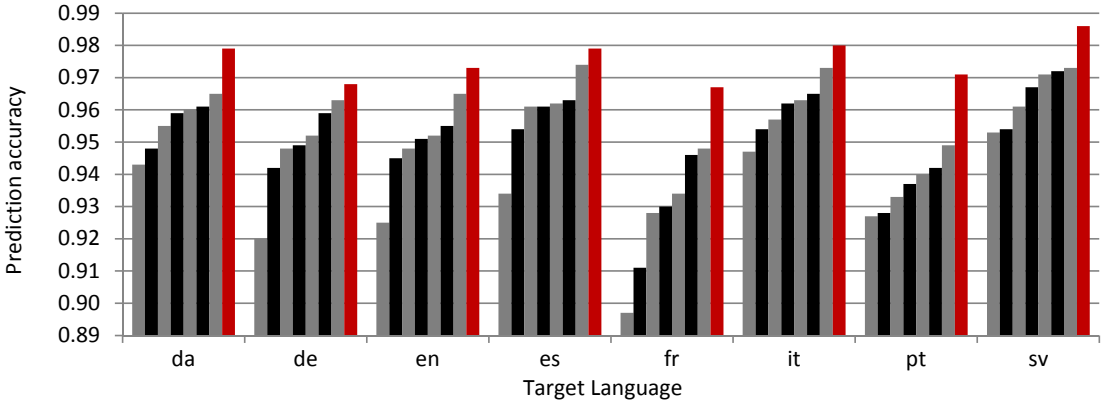

Figure 6.3: Visualisation of Table 6.1. Prediction accuracies of the 7 baseline source languages and the multi-lingual training set, shown for each of the eight target languages. Bars are shown in ascending order of accuracy, with the multi-lingual training set result marked red.

	da	de	en	es	fr	it	pt	sv
Danish (da)	—	+.005	+.013	+.009	+.**033**	+.011	+.013	−.009
German (de)	+.015	—	+.016	+.**031**	+.**035**	+.009	−.002	−.001
English (en)	+.**021**	+.018	—	+.**022**	+.**040**	+.010	+.005	+.010
Spanish (es)	+.005	+.**062**	+.**021**	—	+.**024**	+.017	+.**035**	+.004
French (fr)	+.015	+.016	+.011	+.017	—	+.000	+.018	+.010
Italian (it)	−.003	+.017	+.011	+.**025**	+.019	—	+.010	+.017
Portuguese (pt)	+.**024**	−.001	−.**026**	+.**025**	+.011	+.**022**	—	+.011
Swedish (sv)	+.009	+.011	+.**025**	+.019	+.**061**	+.**030**	+.017	—

Table 6.2: Difference in prediction accuracy after feature selection when compared to the corresponding results in Table 6.1. As in Table 6.1, rows 2-9 denote the training language, columns 2-9 denote the testing language. Differences of more than .02 are highlighted.

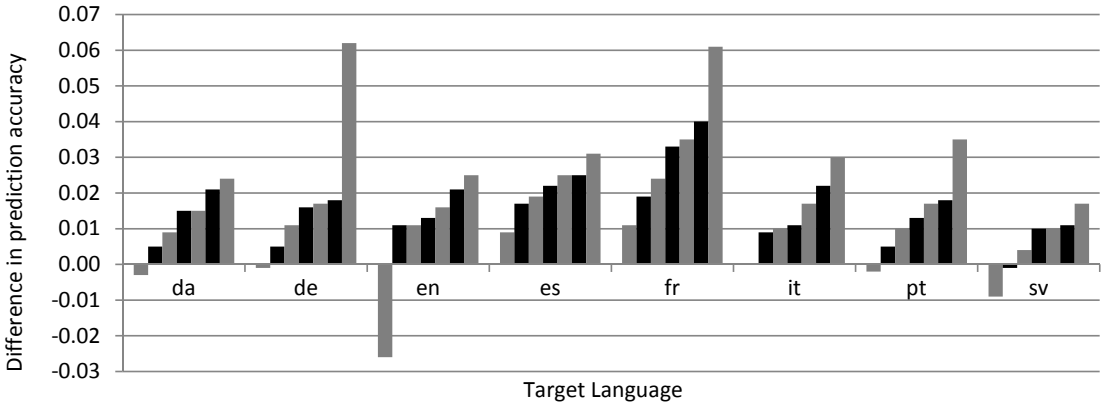

Figure 6.4: Visualisation of Table 6.2. Difference in prediction accuracies between the reduced and the full feature sets. Positive and negative values show that the reduction in features increased and decreased the accuracy, respectively. Results are shown for all 56 combinations of source and target languages. Note that the sizes of the reduced feature sets differ for each combination, as explained in the text. For each target language, bars are sorted in ascending order of accuracy difference.

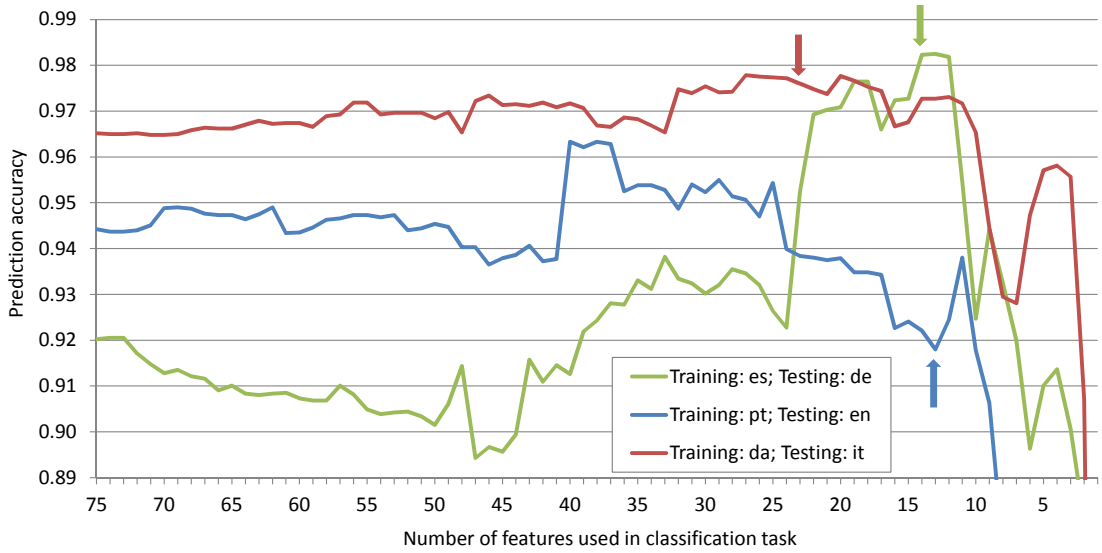

Figure 6.5: Prediction accuracies for the es→de, pt→en, and da→it classification tasks for each possible value of k (number of features). The arrows in the respective colours indicate the values for k chosen by the cross-lingual cross-validation method.

Since these subsets were identified using a supervised ranking technique, the results illustrated in Figure 6.4 suggest that comparable corpora can also be used to identify features with strong discriminative powers for cross-lingual genre classification tasks. They also show that this is possible even if neither the source nor the target language is included in the comparable corpus.

An important question is whether the algorithm can find a good value for the threshold k. Based on the results discussed above, three combinations of source and target languages were picked and observed in more detail. They were the one that gained the most from the feature reduction (training on Spanish texts, testing on German texts: es→de), the one that suffered the most (training on Portuguese texts, testing on English texts: pt→en), and the one that used the largest number of features (training on Danish texts, testing on Italian texts: da→it). For these three combinations, the performance when removing features from the set one by one were recorded, starting at the performance of the full set.

Figure 6.5 illustrates the prediction accuracies as functions of the number of features used. The arrows indicate the threshold chosen by the algorithm. The es→de classifier performs clearly better when selecting between 12 and 22 features from the ranking. The threshold value (14) happens to be a very good choice and yields considerable improvement over the baseline. The performance of the pt→en classifier stays mostly

within the confidence interval of the baseline[1], although it clearly outperforms it for feature set sizes 37-40. Accuracy drops and falls below baseline level for fewer than 20 features. Here, the chosen threshold (13) is too low, since this classifier would benefit from additional features. The da→it classifier benefits slightly but significantly from a reduced feature set until accuracy drops sharply for less than 11 features. The threshold (23) is a good choice, although the exact value is less crucial than for the es→de and pt→en classifiers, in that small variations would have little effect on the result.

The majority of positive results in Figure 6.5 suggests that the chosen threshold k is usually suitable to improve the prediction accuracy. In line with that, Figure 6.5 shows that the algorithm picks a near-optimal value for k for some source-target combinations. However, the example of the pt→en classifier shows that this is not necessarily the case. On the other hand, it also illustrates that even where feature reduction leads to deteriorating performances, this could be due to a sub-optimal threshold choice. This is clearly the case for the pt→en classifier, where a set of 37-40 features would have improved baseline performance significantly. Optimizing the computation of this threshold, possibly by exploiting the unlabelled data in the target language, would be an interesting problem for future work.

In order to get an idea of the types of features which are typically selected, they were ranked by their information gain using a combined set that included all eight languages. The top 15 features are listed below. Note that the information gain of a certain feature varies depending on the exact set of languages used. However, the ranking in the experiments described above was fairly stable and the top 15 features rarely differed from the ones below.

1. `Single sentence paragraph count`
2. `Single sentence paragraph/sentence ratio`
3. `Average paragraph length`
4. `Closing parenthesis frequency`
5. `Opening parenthesis frequency`
6. `Number/token Ratio`
7. `Forward slash frequency`
8. `Single sentence paragraph distribution`
9. `Colon frequency`

[1]This is calculated by assuming that the number of misclassifications is approximately normally distributed with mean $\mu = en$ and standard deviation $\sigma = \sqrt{\mu(1-e)}$, where e is the fraction of misclassified instances and n is the size of the test set. The 95% confidence interval is then $\mu \pm 1.96\sigma$.

10. Average sentence length
11. TF-iDF average precision
12. Type/token ratio
13. Document length
14. Paragraph length standard deviation
15. Hyphen frequency

As expected, none of the 25 common-word frequency features was ranked among the top 15. This finding reinforces the intuition that common-word frequencies are useful in mono-lingual genre classification tasks, but add little to cross-lingual models and may even be harmful. While feature 11 above seems to have discriminative power, none of the other TF-iDF based features is in the above list. This is likely due to the fact that these features have informative value only in combination with each other. However, information gain ranking evaluates only single features, not sets. A subset based selection approach might be more suitable to identify their strengths (cf. Guyon and Elisseeff 2003).

Another observation is that features based on paragraph length dominate the ranking. This is likely due to the way texts of the three different genres are structured. Legal texts tend to have very short paragraphs, sometimes consisting of a single token (Example 1 below). Newswire paragraphs are mostly only one or two sentences long, but typically contain more than one token each (Example 2). In transcribed speech (Example 3), paragraphs tend to be longer.

1. **Legal text:**

```
<p>Commission Regulation (EC) No 1135/2006</p>
<p>of 25 July 2006</p>
<p>amending the import duties in the cereals sector applicable from 26 July
2006</p>
<p>THE COMMISSION OF THE EUROPEAN COMMUNITIES,</p>
<p>Having regard to the Treaty establishing the European Community,</p>
```

2. **Newswire text:**

```
<p>The KFX top-20 index lost 0.20 point to close at 126.29 in overall bourse
turnover of 1.944 billion crowns.  The KFX December future rose 0.65 point
to 126.40 with 10 contracts each worth 100,000 crowns traded.</p>
<p>Novo Nordisk attracted a good deal of attention following its announcement
of 400 million crown rationalisation cuts for 1997 and 1998, finishing the
```

day a solid 21 crowns up at 954.

3. **Transcribed speech text:**

Naturally I understand the honourable Member's concern. As far as the Commission is concerned, we have never supported financially the production or distribution of school textbooks nor the preparation of school curricula. Assistance to the educational system is focused mainly on infrastructure, equipment for schools and direct assistance for school expenses, for example, salaries. No request has ever been made by the Palestinian Authority to the Commission to finance school curricula and textbooks.

Chapter 7

Label Propagation for Cross-Lingual Genre Classification

As the results presented in Chapter 5 show, exploiting unlabelled texts written in the target language can boost the performance of a cross-lingual genre classifier. Improvement comes from using additional text features with predictive power for this task, which are specific to the target language and so cannot be used across languages. While the iterative re-labelling algorithm achieves this for broad genre categories and a small set of target classes, it is vulnerable to over-predicting dominant genre classes in more fine-grained settings. Furthermore, during training of the SVM model on the source language texts, information from the target language is not considered. Therefore, features which separate genre classes well in the source language are assigned a high weight, which hurts classifier performance if these features do not separate genres in a similar way in the target language. As the iterative re-labelling algorithm relies on a good initial classification of at least some instances, target language adaptation cannot remedy a poor cross-lingual labelling, and might even downgrade results. A prior selection of features can benefit accuracy by identifying features that work well cross-lingually, as demonstrated in Chapter 6. However, that approach requires labelled texts in several languages, which may not be available for the specific genre classification task.

In order to take better advantage of unlabelled target language texts, a graph-based classifier based on the label propagation algorithm of Zhu and Ghahramani (2002) is proposed in this chapter. Section 7.1 gives an overview of how this method addresses the aforementioned shortcomings of the iterative re-labelling approach. It briefly reviews the label propagation algorithm of Zhu and Ghahramani (2002) and describes task-specific extensions. Section 7.2 describes the experiments carried out to evaluate the

method, while Section 7.3 discusses the results.

7.1 Graph-based learning

7.1.1 Basic Algorithm

The label propagation algorithm proposed by Zhu and Ghahramani (2002) was designed as a semi-supervised solution to problems where both labelled and unlabelled data are available. While it is not specifically targeted at (cross-lingual) text classification, it fits the requirements for a method that can exploit unlabelled data and (with an extension) is well suited to combining different sets of features. The algorithm described in this sub-section is exactly as proposed by Zhu and Ghahramani (2002), but further detail, evaluation, and extensions can be found in the original publication.

Before the actual propagation process begins, a fully connected graph is constructed, where all $|N|$ (labelled and unlabelled) data instances are represented as nodes. Edges between nodes are assigned weights based on the Euclidean distances between the corresponding data points:

$$w_{ij} = w_{ji} = \exp\left(-\frac{\sum_{d=1}^{D}(x_i^d - x_j^d)^2}{\sigma^2}\right)$$

Here D is the total number of dimensions, or features, and σ is a parameter which Zhu and Ghahramani (2002) set by a heuristic (see Section 7.2.2). Based on these weights, a transition matrix T is created in which the weights are column-normalized to ensure that each node has the same total output weight during the propagation process (although note that nodes may have differing total input weights):

$$T_{ij} = \frac{w_{ij}}{\sum_{n=1}^{N} w_{nj}}$$

In addition, each data point is assigned a vector of class probabilities. For labelled instances, vector elements are initialized to 1 for the class label, and 0 for all other classes. Zhu and Ghahramani (2002) show that the initial values for unlabelled data points are irrelevant. All class probability vectors are represented in a $|N| \times |C|$ matrix Y, where $|C|$ is the number of classes.

In each iteration of the algorithm, the class labels are propagated along the edges of the graph simultaneously:

$$Y \leftarrow TY$$

Y is then row-normalized, so that the values in each row can be interpreted as class probabilities. The rows corresponding to labelled instances are clamped to their original binary state. The propagation, normalization, and clamping steps are repeated until convergence.

The final classification is derived by assigning each unlabelled instance to the class with the highest probability in the corresponding row of Y. In order to prevent over-prediction of dominant classes, Zhu and Ghahramani (2002) suggest scaling the class probabilities so that the column sums of Y fit an expected distribution (e.g., that of the labelled instances). They also propose an alternative label bidding process, where the number of instances for each class is given and instances bid for class labels using their label probabilities.

7.1.2 Multi-Layered Graph

In order to adapt Zhu and Ghahramani's label propagation algorithm to the task of CLGC, several adjustments are made. The most important is an extension involving the construction of a multi-layered graph with genre-specific edge weights. Here, each layer of the graph corresponds to one set of features. These layers share nodes, but the edges between the nodes are specific to each layer. One of the reasons for this extension is that it allows different features to be used to assign edge weights between target language nodes than those used for cross-lingual edges. As the results of Chapter 5 have shown, this can improve classification results.

However, a multi-layered graph allows for more than two separate feature sets. In theory, it supports infinitely many layers, both cross-lingual and mono-lingual. In this project, the distinction between features is mostly motivated by different levels of required resources (Chapter 3). Feature sets were therefore not specifically designed to correlate with particular characteristics, or facets, of a text. However, as the types of features they contain differ strongly from set to set, it is expected that this separation of layers, as well as the weighting algorithm introduced in this section, will benefit the classifier.

With enough knowledge of the target language, feature sets based on genre facets (see Section 1.1) are also possible. Assume, for example, that prior work had established two given sets of features to be good predictors for the level of subjectivity and the level of complexity, respectively. They could be used to construct one layer of the graph each. Because of the genre-specific edge weights introduced in this section, the algorithm

then would have the potential to identify which genre is described well by which facet, both across languages and within the target language. Labels could then be mostly propagated through the corresponding layer, while the layer corresponding to the less descriptive facet for a genre is used less. Note however that an evaluation of facet-based feature sets is left for future work.

In order to construct a multi-layered graph, the Euclidean distances between data points (i.e, texts) in each feature set $f \in F$ are computed and weights are assigned using a Gaussian function, as before:

$$w_{ij}^{f} = w_{ji}^{f} = \exp\left(-\frac{\sum_{d_f=1}^{D_f}(x_i^{d_f} - x_j^{d_f})^2}{\sigma^2}\right)$$

The algorithm described by Zhu and Ghahramani (2002) assumes that all data points can be represented by the same set of features. However, as explained in Chapter 3, for CLGC there are features that may be useful within the target language but cannot be used across languages. This means that we cannot compute distances between source and target language texts in the mono-lingual feature sets. Therefore, in the mono-lingual feature sets, distances are only computed among target language texts. In the cross-lingual feature sets, distances are computed between source and target language texts, but not between texts of the same language.

Each graph layer is assigned a separate transition matrix T^f. These are computed in the same way as before. Since distances are not computed for all pairs of texts, the values in T^f without a corresponding w_{ij}^f value are set to zero.

One advantage of using multiple layers in the graph is that the impact of each feature set is equal, regardless of its dimensionality or how the feature values are scaled. This can be beneficial in the first iteration, as it is unknown which feature sets are helpful in identifying the different genres. To this end, class labels are propagated through the edges of the cross-lingual graph layers, and the resulting target text class probabilities are summed over the feature sets:

$$Y \leftarrow \sum_{f=1}^{F} T^f Y$$

Because of the subsequent row-normalization of Y, this effectively averages the inputs a given node receives over all graph layers.

However, the structure of such a multi-layered graph also allows for easy adjustments of the edge weights during the label propagation process. This facilitates an implementation based on the intuition that different genres are identified by different characteristics.

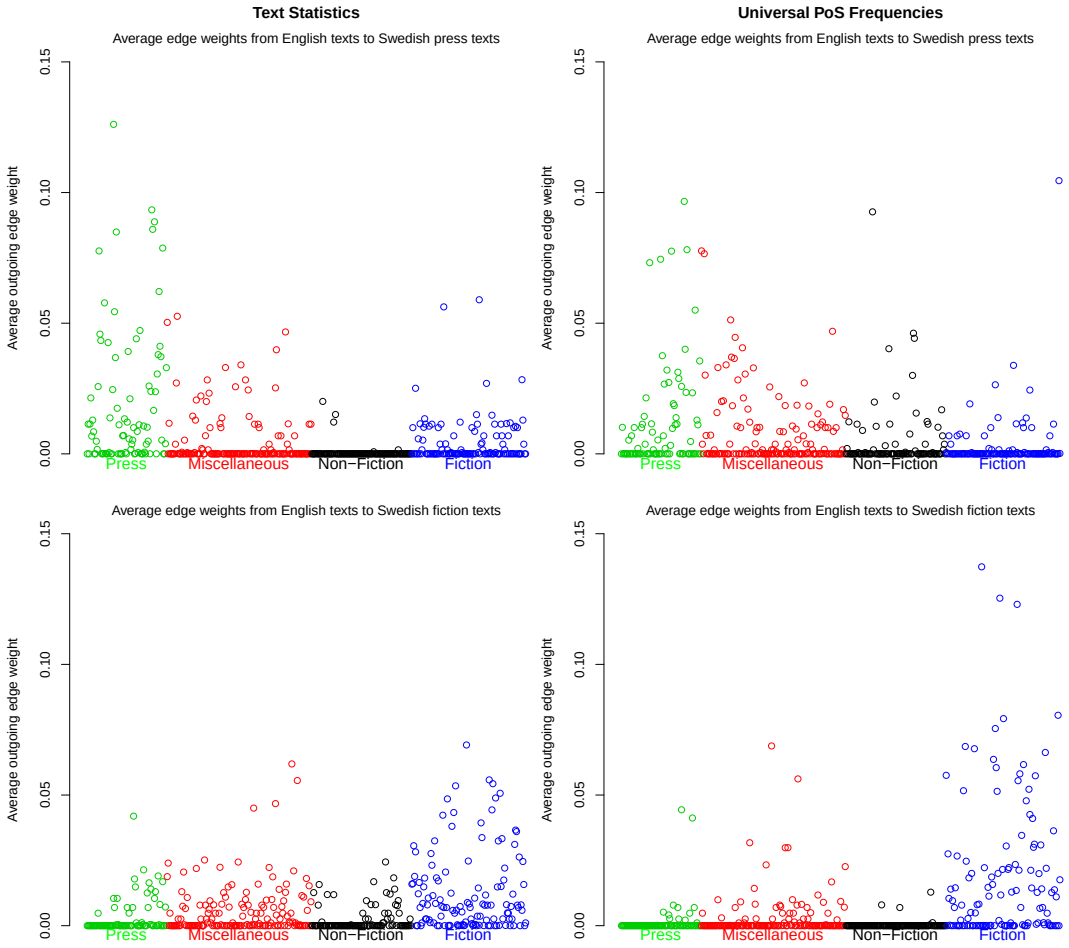

Figure 7.1: Plot of the average edge weights from the 483 nodes representing English texts in the text statistics (left) and universal PoS (right) layers. Each dot represents one node, with colours indicating the genre of the corresponding text. Averages are computed over all nodes representing Swedish press (upper) and fiction (lower) texts.

To illustrate that two genres can be more reliably described by different types of features, consider Figure 7.1. It assumes English as source and Swedish as target language, and plots the outgoing cross-lingual edge weight averages of nodes corresponding to English texts. The averages are computed over Swedish press and fiction texts respectively and for two layers of the graph, corresponding to text statistics (TS) and universal PoS frequencies. In other words, the higher a value for a given node in the plot, the larger is its impact on the Swedish press/fiction texts. Therefore, high values are desirable for the English press texts (green) in the upper plots, and for the English fiction texts (blue) in the lower plots. Figure 7.1 reveals the genre-specific differences between the graph layers. While in the TS feature set the nodes with the highest impact on Swedish press nodes correspond to English press texts, this is not the case for the universal PoS frequencies. Swedish fiction texts, on the other hand, receive more impact from English fiction nodes through the PoS-based layer, while the input from the TS layer is more balanced across genres. Of all incoming edge weight to Swedish press texts, 34% and 56% comes from English press texts through the TS and PoS layers respectively. For Swedish fiction texts, the percentages are 79% and 53%, respectively. It would therefore be preferable if press nodes propagate their labels through a different layer than fiction nodes.

Note that the plots in Figure 7.1 require knowledge of target language genre labels, which is of course unavailable to the classifier. They can however be estimated from the Y matrix. After each iteration, label product matrices L^c are computed for each genre class $c \in C$:

$$L^c = Y_c Y_c^\top$$

where Y_c is the c-th column of Y. Each entry in L^c represent the probability of two texts belonging to the same genre class c. High values can only occur if both texts have a high probability of belonging to this genre.

A visual example is shown in Figure 7.2. The high probabilities of Node 1 and Node 2 belonging to Genre A result in relatively high values of $L^A_{1,2}$ and $L^A_{2,1}$. Similarly, Node 3 and Node 4 are likely to belong to Genre B. Therefore $L^B_{3,4}$ and $L^A_{4,3}$ are the fields with the highest values in L^B.

The L^c matrix can then be used to compute modifier values for each feature set and each genre class, which are stored in a row-normalized $|C| \times |F|$ modification matrix M:

$$M'_{cf} = \sum_i \sum_j T^f_{ij} L^c_{ij}$$

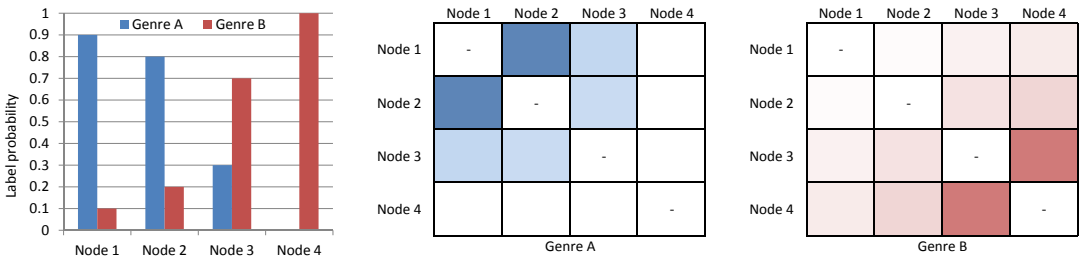

Figure 7.2: Left: Example class probabilities stored in Y for four texts (nodes) and two genres. Center and Right: Resulting label product matrices L^A and L^B. Darker fields indicate higher values.

$$M_{cf} = \frac{M'_{cf}}{\sum_{k=1}^{F} M'_{ck}}$$

Each row of M represents a distribution over all feature sets for one of the genre classes. This distribution describes how well the weights in the corresponding graph layers represent the current labelling for this genre class. In other words, a feature set f that, on average, puts texts with a high probability of belonging to a certain genre c closer to each other than competing feature sets, will result in a high value in the corresponding field M_{cf}.

Note that M_{cf} can be derived directly, without first computing L^c, because

$$\begin{aligned} M'_{cf} &= \sum_i \sum_j T^f_{ij} L^c_{ij} \\ &= \mathrm{tr}(T^f L^{c\top}) \\ &= \mathrm{tr}(T^f Y_c Y_c^\top) \\ &= \mathrm{tr}(Y_c^\top T^f Y_c) \\ &= Y_c^\top T^f Y_c \end{aligned}$$

The edge weights of each layer can then be adjusted depending on the belief about which genre class a node belongs to. First, we compute a $|N| \times |F|$ matrix R, containing the modification values for each node and each feature set depending on the class distribution of the node:

$$R = YM$$

An example for this process is shown in Figure 7.3. The edge weights in the two layers X and Y are a reasonably good fit for the label product matrices L^B and L^A shown in Figure 7.2 respectively. Therefore, $R_{1,2} > R_{1,1}$ and $R_{2,1} > R_{2,2}$. In other words, nodes

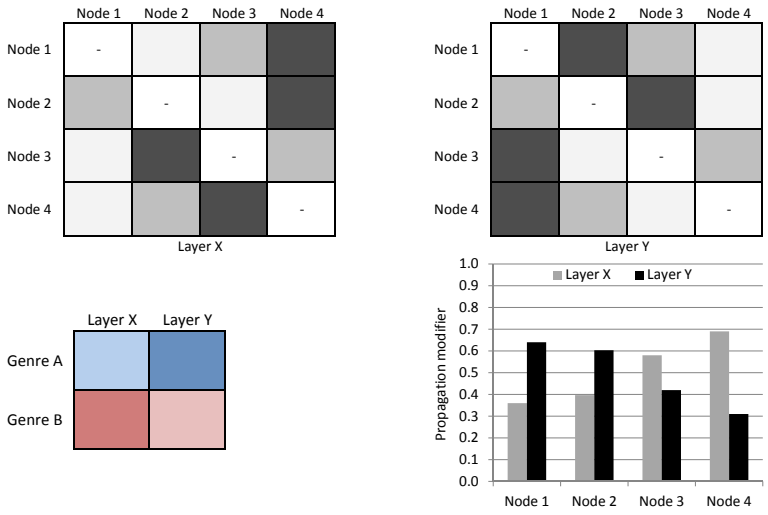

Figure 7.3: Top: Example edge weights for two layers X and Y. Rows denote receiving nodes, columns denote propagating nodes. Darker fields denote higher values. Bottom left: Genre-specific layer weights stored in M, based on the two layers and the matrices L^A and L^B shown in Figure 7.2. Bottom right: Modification value stored in matrix R for each node based on its label probability.

with high probabilities of belonging to Genre A will propagate their labels more through Layer Y than Layer X. The opposite is true for nodes with high probability of belonging to Genre B. This is shown in the bottom right graph of Figure 7.3.

This modification of edge weights is implemented by computing adjusted transition matrices T'^f which are then used to propagate labels in this iteration:

$$T'^f_{ij} = T^f_{ij} R_{jf}$$

Note that by multiplying T^f_{ij} with R_{jf} instead of R_{if}, the class probability of the propagating node is considered, but not that of the receiving node. This means that texts which have a high probability of belonging to a certain genre will communicate this belief mostly through the graph layers strongly associated with that genre. This high probability does not affect how the corresponding node receives input however.

The actual label propagation step is the same as before, except that T^f is replaced by T'^f:

$$Y \leftarrow \sum_{f=1}^{F} T'^f Y$$

As before, the label matrix Y is row-normalized and the labelled nodes (i.e. source language texts) are clamped to their initial state.

7.1.3 Rank-Based Weights

Another adjustment to Zhu and Ghahramani's algorithm involves the computation of edge weights based on ranks, rather than absolute distances. (The motivation for this will be discussed shortly). A distance ranking in ascending order is computed for each target language text in each feature set. This then serves as a base for calculating edge weights, rather than using the absolute Euclidean distances. The weights are still computed by a Gaussian function:

$$w_{ij}^f = \exp\left(-\frac{(r_{ij}^f - 1)^2}{\sigma^2}\right)$$

where r_{ij}^f is the rank of node j in the distance ranking of node i for feature set f. Note that r_{ij}^f is typically different from r_{ji}^f if $i \neq j$ and therefore so are w_{ij}^f and w_{ji}^f (unlike in the original label propagation algorithm). Since w_{ij}^f is the weight of the directed edge from node j to node i, the rank-based weight calculation ensures that all nodes receive the same total input within a layer of the graph. To avoid losing this advantage, the transition matrices T^f are not column-normalized:

$$T_{ij}^f = w_{ij}^f$$

Again, any edge for which no weights are computed (because the corresponding nodes cannot be represented in the same feature space or because the edge is not needed) is represented by zeros in the transition matrices.

There are several advantages of using ranks instead of absolute distances for edge weight calculation. Firstly, it makes the choice for the parameter σ easier. One reason for this is that only two different values are needed for σ: One for all cross-lingual feature sets, and one for all mono-lingual feature sets. This is because the rank values are identical and σ should therefore depend solely on the length of the ranking, which is $|N_S|$ for the cross-lingual, and $|N_T|$ for the mono-lingual feature sets. With weights based on absolute values, on the other hand, each feature set would require a sensible value for this parameter based on the structure of the data.

Another advantage is that each node is guaranteed the same total input weight within a layer of the graph but nodes can have very different total output weights. This ensures that outliers receive labels from several sources, but have little or no impact on other nodes. Furthermore, each layer will have the exact same total impact on the label probabilities assigned to a node, as long as σ is identical. This means that all cross-lingual feature sets have the same weight, as do all mono-lingual feature sets.

Since different sets can vary in dimensionality and scale, this ensures that smaller sets are not under-represented.

7.1.4 Rewarding high label confidence

One element that proved very beneficial experimentally in the target language adaptation process of the iterative re-labelling algorithm (Chapter 5) was the computation of a label confidence score as a basis for the selection of instances for re-training the SVM model, as shown in Figure 5.7. While there it is implemented by excluding texts with low confidence in their predicted labels, the class probability vectors of the label propagation algorithm allow for a more elegant solution. The more even the distribution of genres is for a given text, the lower the confidence in its labels. As these labels (stored in matrix Y) are what is propagated through the graph, a lower output weight for the corresponding node would be desirable. Therefore, a confidence value c_j is computed for each node j:

$$c_j = \sum_c Y_{jc}^2$$

This is then used to modify the outgoing edge weights for each node and each graph layer:

$$T_{ij}'^f \leftarrow T_{ij}'^f c_j$$

Note that this does not have any impact on edges from source language texts. Since the corresponding rows in Y are binary, c_j for these nodes will always be 1.

7.1.5 Constant vs. decreasing source input

Two variants of the label propagation method were implemented. The first one uses a constant input through the source-target edges, that is no edge weight adjustments are made during the iterative process beyond those of the genre-specific layer weights (Section 7.1.2) and label confidence (Section 7.1.4). The second variant follows the intuition that the structure of the target language texts requires less and less input from the labelled source texts. That is, while in the early iterations, a strong input from nodes with clamped label probabilities is helpful to avoid over-prediction of dominant genres, it is less crucial later on, when labels have propagated throughout the graph and genre-specific edge weights have been established among target nodes. On the other hand, this source input can inhibit positive developments of label probabilities due to differences in the source and target language. The idea is therefore to initially

have strong source input, which is decreased in later iterations in order to enable the algorithm to fully exploit the structure of the unlabelled target language set. Note that Zhu and Ghahramani (2002) proposed constant source input, as they assumed that labelled and unlabelled data came from the same distribution, which is obviously not the case in a cross-lingual task.

The decreasing source input was implemented as follows. First, weight adjustment values $\alpha_a = \frac{1}{a}$ and $\beta_a = 1 - \alpha_a$ are computed for the current iteration a. Then, the source-target edge weights are reduced:

$$T'^{f}_{ij} \leftarrow T'^{f}_{ij} \alpha_a$$

where i and j correspond to target and source language texts, respectively. In the next step, the total removed input weight of each target node i is computed:

$$\delta^f_i = \frac{\beta_a}{\alpha_a} \sum_j T'^{f}_{ij}$$

This is then used to determine an edge weight for self loops of target nodes:

$$T'^{f}_{ii} \leftarrow \delta^f_i \beta_a$$

This self loop is required in order to keep the algorithm converging without the constant source input. It effectively replaces some of the belief propagated from the source nodes by a node's own beliefs from the last iteration. Note that the β_a factor means that more of the the removed input weight is used for self loops in later iterations than in earlier ones.

7.1.6 Predicting the genre of new texts

The label propagation algorithm is inherently transductive – all data points that are to be labelled need to be included in the graph. This means that new texts, which are not available at training time, cannot be assigned a genre, unless the whole propagation process is repeated. For offline applications, such as classifying genres in a large corpus, this does not pose a problem. However, it is not practical in online applications such as web crawling, where the genre should be predicted quickly when a text is discovered.

Fortunately, this can be remedied by using the newly labelled texts in the target language as a training set for a separate inductive classifier, which can then be used to predict the genres of new texts. An obvious choice here would be instance based methods, such as k-nearest-neighbours, since they could make use of the distribution

over labels assigned to data points. However, they require substantial computation at test-time, since the k nearest nodes have to be identified. This contradicts the requirement for a fast online prediction. It therefore makes more sense to use a pre-trained, inductive classification model, such as an SVM, for this task. To this end, the target language texts are assigned hard genre labels after the label propagation algorithm has converged, as described in Section 7.1.1. They can then be used as a training set with the preferred feature representation for the task and language at hand.

The advantage over training on source language texts is that mono-lingual features can be used, which are less restricted, more adaptable to the target language, and better researched. Furthermore, the output of the label propagation algorithm allows for a straightforward selection of texts to train the subsequent classifier on. Provided enough data is available, it is possible to exclude texts with little confidence in their label – that is, a flat distribution in the corresponding row of Y.

To this end, a threshold parameter t is used with $0 < t \leq 1$. This determines the percentage of texts from each genre that are used for training. The absolute number of selected texts for a genre c is then $s_c = tp_c$, where p_c is the number of texts assigned to c. This means that the texts in the training set $S = \{s_1, ..., s_C\}$ will be distributed across genres proportionally to the outcome of the label propagation algorithm, though not necessarily proportionally to the source language texts.

7.1.7 Complexity

The time and space requirements of the method as described above are functions of the number of texts N, the number of features used M, the number of feature sets F, and the number of genre classes C. Here, the number of features can be described as $M = Fm$, where m is the average number of features per set. Furthermore, the number of iterations I is a factor. While the time of convergence may depend on the previously mentioned variables, I treated as a separate input variable for complexity calculation.

Before the iterative algorithm starts, the Euclidean distance is computed for each pair of texts, in each feature set. This computation requires quadratic time w.r.t. the numbers of texts, that is $O(MN^2) = O(FmN^2)$. However, there is no need to compute distances among source language texts. This is because source language texts will not propagate labels between themselves (their label probabilities are clamped) and distances are not required to determine a value for σ, due to the rank-based edge weights (see Section 7.2.2).

Therefore, N can be split into the number of source and target language texts, $N = N_S + N_T$. Similarly, F can be split into the numbers of cross-lingual features and target language specific features, $F = F_{ST} + F_T$, with the average number of features described by m_{ST} and m_T respectively. By not computing source to source distances, complexity can be reduced to $O(F_{ST} m_{ST} N_T N_S + F_T m_T N_T^2)$, that is the computation time is quadratic w.r.t. to the number of target languages texts, but linear w.r.t. the number of source language texts.

As edge weights are rank-based, rather than distance-based, values have to be sorted based on the distance to a given target language text. As this is done for each target language text, this requires $O(F_{ST} N_T N_S \log(N_S) + F_T N_T^2 \log(N_T))$, assuming a sorting algorithm with $O(N \log(N))$ complexity, such as quicksort or timsort.

The actual label propagation is a multiplication of two $N \times N$ and $N \times C$ matrices, done for all F feature sets and in each of the I iterations, that is $O(IFCN^2)$. However, as mentioned, labels do not need to be propagated from source to source texts. Therefore, the same can be achieved by multiplying a $N_T \times N_S$ with a $N_S \times C$ matrix and a $N_T \times N_T$ with a $N_T \times C$ matrix. This means that the time complexity of this step is $O(ICF_{ST} N_T N_S + ICF_T N_T^2)$. All other steps during in the iterative process require equal or less time w.r.t. these variables.

The algorithm as described in this chapter therefore runs in linear time w.r.t. I, F, m, and C. It requires $O(N_S \log(N_S))$ w.r.t. N_S and $O(N_T^2 \log(N_T))$ w.r.t. N_T.

The space requirements depend on the size of the matrices. There are F $N \times N$ matrices containing edge weights. However, these can be represented as F_{ST} $N_S \times N_T$ matrices and F_T $N_T \times N_T$ matrices. Furthermore, the genre label probabilities of each node are stored in a $N \times C$ matrix and the computation of genre-specific weights involves two $C \times F$ and $N \times F$ matrices. The overall space complexity of $O(F_{ST} N_T N_S + F_T N_T^2 + NC + CF + NF)$ means that space requirements are linear w.r.t. all variables but N_T, for which they are quadratic, and I, for which they are constant.

However, the algorithm can be optimized by ignoring the long tail of the Gaussian function and keeping only the values above a threshold for the minimum edge weight in the transformation matrices T^f. As ranks are used for assigning edge weights, this is equivalent to choosing the top k ranks. This means that a selection algorithm can be used to find the k^{th} node in the ranking, such as quickselect, which has linear time complexity. All distances larger than the one at rank k can be ignored and what remains is a sort problem with k elements, requiring $O(k \log(k))$. As k is dependent on the parameter σ, which in turn increases with $\log(N_S)$ or $\log(N_T)$ (see Section 7.2.2), this

requires only $O(\log(N_T)\log(\log(N_T)))$ or $O(\log(N_S)\log(\log(N_S)))$, which is less than linear and therefore less than the selection algorithm. During the iterative process, complexity is also reduced, as the multiplication of the now sparse matrix T^f with Y only requires $O(CF_{ST}N_T\log(N_S)+CF_TN_T\log(N_T))$ in each iteration. Therefore, overall time complexity can be described as

$$O(F_{ST}m_{ST}N_TN_S+F_Tm_TN_T^2+ICF_{ST}N_T\log(N_S)+ICF_TN_T\log(N_T))$$

The optimized algorithm then runs in linear time w.r.t. I, F, m, C, and N_S, but quadratic time w.r.t. N_T.

Space requirements are also reduced by this optimization, as only an $N\times k$ matrix is required to store edge weights. This results in a space complexity of $O(F_{ST}\log(N_T)N_S+F_TN_T\log(N_T)+NC+CF+NF)$, that is the algorithm's space requirements are reduced from quadratic to linearithmic w.r.t. N_T.

Note that the results presented in Section 7.2 were yielded with dense matrices, that is without the optimization techniques described here.

7.2 Experiments and Results

The two variants of the label propagation method (constant and decreasing source input) were evaluated and compared with the iterative re-labelling algorithm and other baselines. In addition, systematic experiments were carried out to determine if and how the adjustments mentioned in Section 7.1 can benefit a cross-lingual genre classifier. This section explains the experimental setup and presents the obtained results, which are then discussed in Section 7.3.

7.2.1 Framework

To compare the performance of the classifier, the experiments described in Section 5.3 were repeated. That is, the two label propagation variants were run on data from the BC, LCMC, and SUC (Section 2.1). The feature sets used here were text statistics (Section 3.1, both for cross-lingual and mono-lingual graph layers) and target language word frequencies (Section 3.4, mono-lingual layer only). The benefit of exploiting the additional knowledge provided by a PoS tagger was also tested. To this end, a cross-lingual layer based on univeral PoS tags, and a mono-lingual layer based on PoS histograms (Section 3.3) were added to the graph. The results of these experiments are reported in Section 7.2.3.

Further goals of the experiments were to evaluate whether the approach also performs well for other genre palettes and language pairs, and which of the extensions to the original algorithm of Zhu and Ghahramani (2002) actually yield improvements. To this end, the NYTAC and TüPP-D/Z corpora of newspaper texts in English and German (Section 2.2) were used for further tests, with English as the source language and German as the target language.

As explained in Section 7.1.2, the label propagation algorithm works with several layers corresponding to different feature sets. In the experimental setup, four cross-lingual and four mono-lingual sets of text features were used. The four cross-lingual feature sets were text statistics, punctuation frequencies, universal PoS frequencies, and translated word frequencies. The first two of these sets were also used mono-lingually, together with two more sets: PoS histograms and target language word frequencies. All of these sets are described in detail in Chaper 2. All feature values were scaled to zero mean and unit variance across all texts of the same language.

In addition, a proof-of-concept experiment was carried out using data from the BNC and CIIL corpora (Section 2.4) to evaluate the classifier on poorly-resourced languages. Here, English was used as the source language and Tamil and Malayalam were used as target languages.

7.2.2 Parameters

The assignment of edge weights requires the parameter σ. Finding a good value for this is important, as Zhu and Ghahramani (2002) show that both too small and too large values result in a poor classification accuracy. As mentioned in Section 7.1.3, the choice is more difficult if weights depend on absolute distances, rather than ranks. For this case, the heuristic suggested by Zhu and Ghahramani (2002) was followed. This involves computing the distance between each pair of labelled data instances (i.e., source language texts). The shortest distance δ between two texts from different genres is then used as a basis to calculate the parameter value for $\sigma = \delta/3$. With multi-layered graphs, δ will differ for each feature set, and therefore each layer has a different value for σ.

With rank-based distances, on the other hand, σ can be determined independently of the actual distances between texts. It should however depend on the number of texts in the source and target languages and the sizes of genre categories. Intuitively, σ should be larger for a problem with two classes of equal size than for a problem with many

small classes. For the experiments, a heuristic is used to compute $\sigma = \ln(\theta)$, where θ is the number of source language texts in the smallest genre class. This is then used for all cross-lingual layers. For mono-lingual layers in the target language, where the smallest genre class is unknown, this is estimated by scaling to the size of the target language set:

$$\sigma = \ln\left(\frac{|D_T| \cdot \theta}{|D_S|}\right)$$

For all variants, the suggestion of Zhu and Ghahramani (2002) was followed and the column sums of the final label probabilities for target language texts in Y were scaled to match those in the source language. For the experiments with Tamil and Malayalam texts, an alternative scaling is evaluated. This requires the input of an oracle to reveal the genre distributions in the target language.

In order to evaluate the inductive SVM extension explained in Section 7.1.6, the threshold parameter t was set to 0.5, that is for each genre, half of the target language texts were used to train the SVM model. The NYTAC and TüPP-D/Z corpora were used for these experiments. As an equal genre distribution in the latter only allowed for 1,308 texts to be used (see Section 2.2), a 36-fold cross-validation was employed to assess the classifier performance. In each fold, 6 different texts of each genre were excluded from the label propagation process and used as a test set to evaluate the SVM model. This ensured that there was enough data for the label propagation (1,272 target nodes), while also giving a reliable result based on 1,296 texts.

7.2.3 Comparative Evaluation

Figures 7.4, 7.5, and 7.6 show the results achieved by the label propagation methods (with constant and decreasing source input) for the two, four, and nine genre tasks respectively. They, like the iterative re-labelling (IRL) approach also shown in these graphs, use only text statistics as cross-lingual features. Two mono-lingual layers are used, based on text statistics and target language word frequencies. Note that this makes only minimal use of the potential strengths of the method described in this chapter, as no cross-lingual genre-specific layer weights can be learned.

The results show that the method often benefits from the iteratively decreasing source input. This means that it can exploit the structure of the unlabelled target language data better if it is less restricted by the input through the cross-lingual edges. However, in some of the nine genre tasks (e.g. en→zh) the opposite can be observed. Here, the iterative process benefits from more regulation from the labelled source

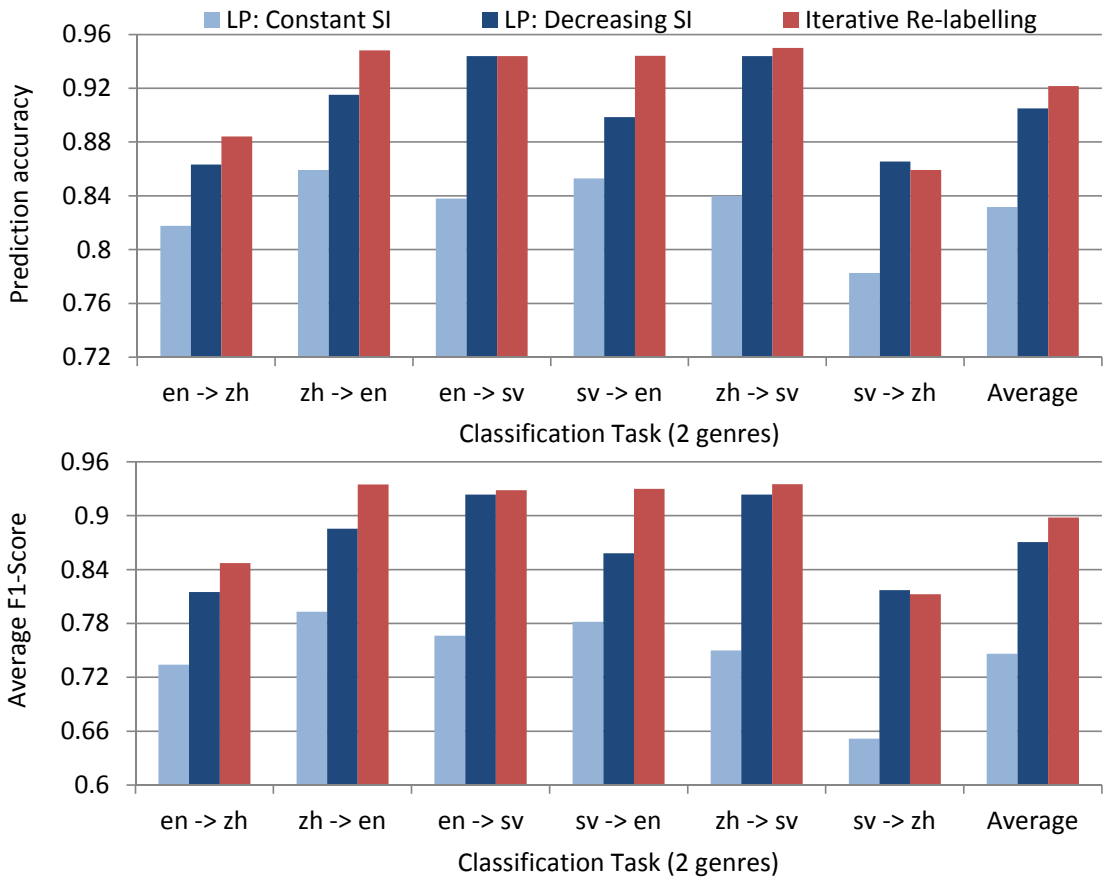

Figure 7.4: Prediction accuracies (top) and average F1-Scores (bottom) for the six classification tasks with two target genres using English (BC), Chinese (LCMC), and Swedish (SUC) texts. The bars indicate the performances of the label propagation method using only text statistics and target language words as features with constant (light blue) and decreasing (dark blue) source input, as well as that of the iterative re-labelling method (red).

language texts, in order to avoid convergence on a suboptimal solution. For most tasks with two or four genres, the decreasing source input variant performs similarly to the IRL method. However, the results in the nine genre tasks reveal a crucial difference. While the IRL classifier achieves a better accuracy for most tasks, this does not translate to superior F1-Scores. The label propagation method, on the other hand, achieves strong F1-Scores whenever it achieves a strong prediction accuracy. This indicates that the latter method is less prone to over-prediction of dominant genres and therefore potentially more suitable for fine-grained and/or imbalanced tasks.

As shown in Section 5.3, the IRL method cannot make much use of an increased

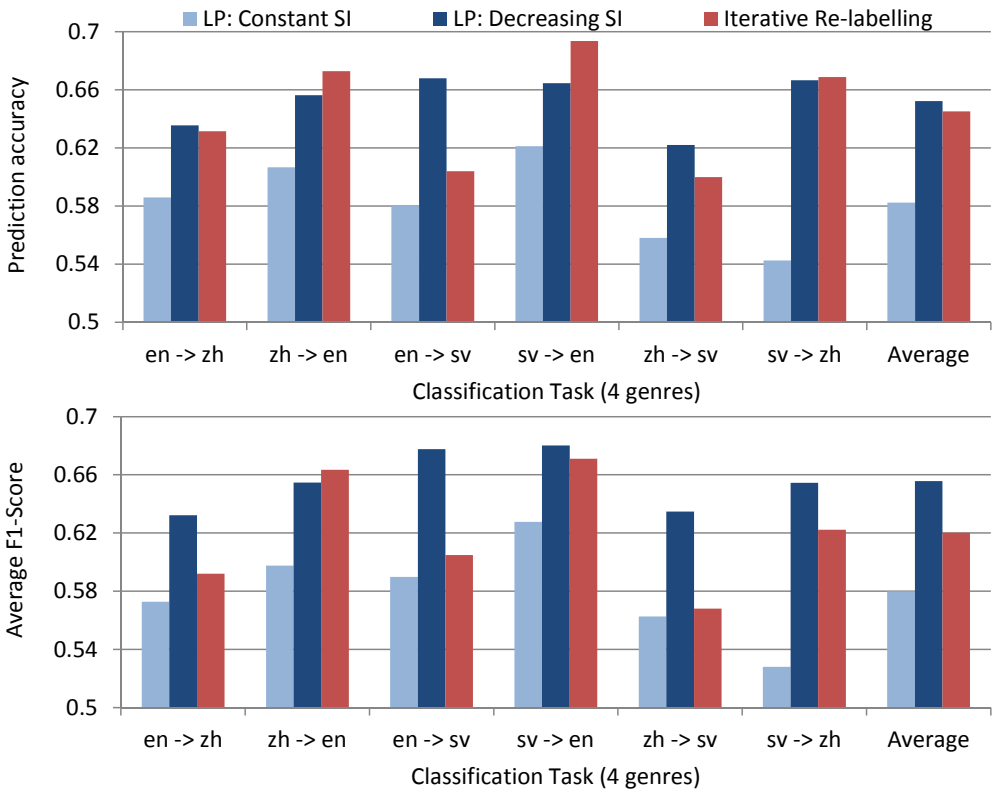

Figure 7.5: Same as Figure 7.4, but for the four genre tasks.

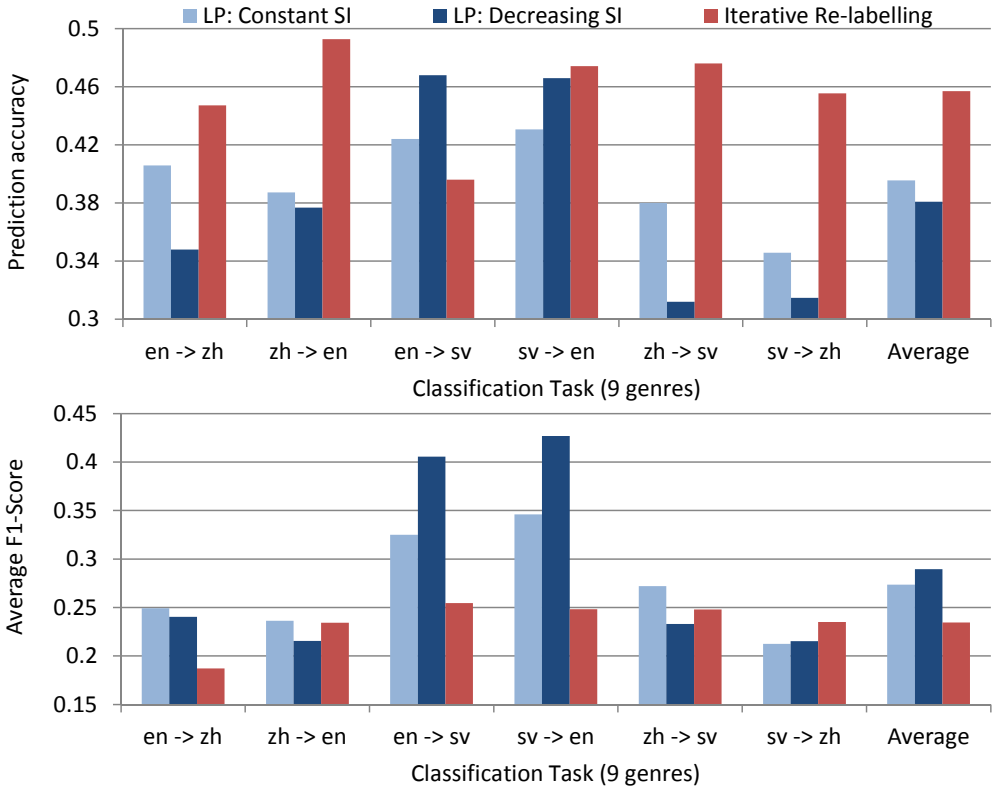

Figure 7.6: Same as Figure 7.4, but for the nine genre tasks.

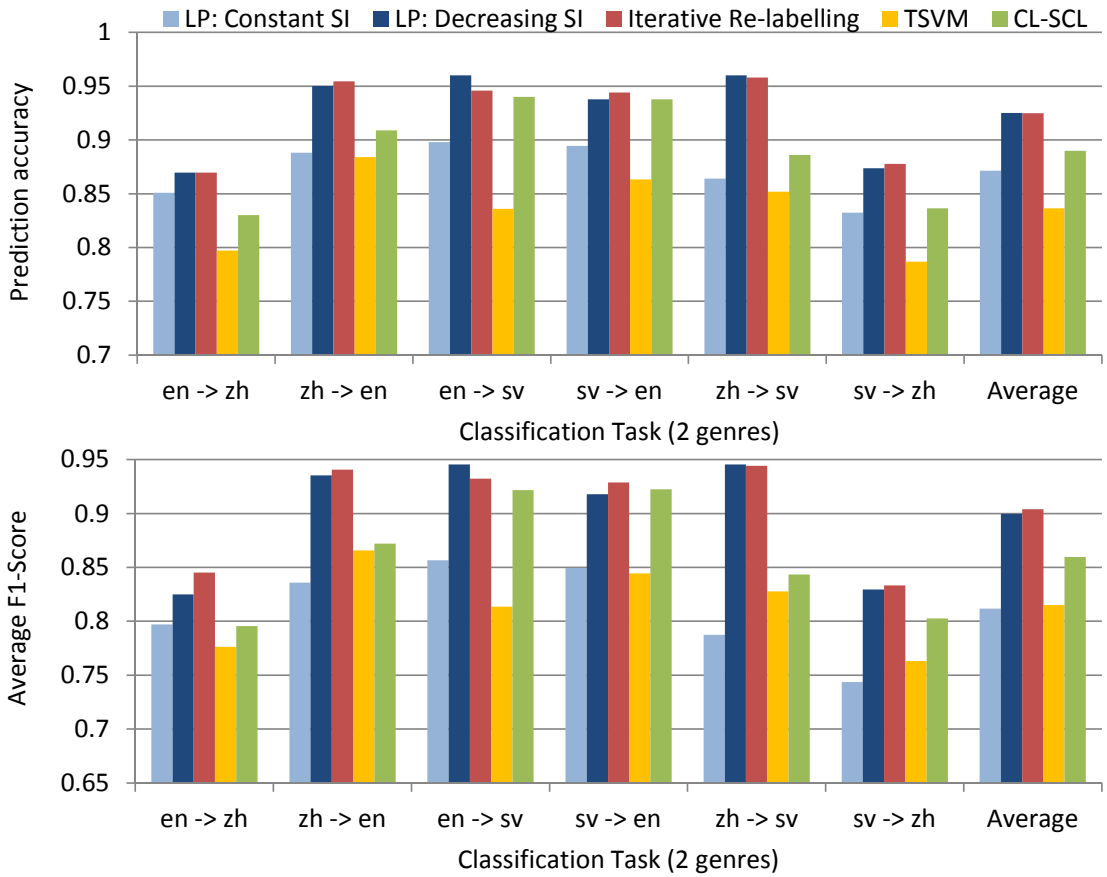

Figure 7.7: Same as Figure 7.4, but with added universal PoS frequencies and PoS histograms as features for the label propagation method and more baselines shown. All methods (except CL-SCL) use the same cross-lingual features.

cross-lingual feature set, should additional resources become available. This scenario was repeated in the experiments in order to evaluate whether this is also the case for the label propagation algorithm. To this end, a cross-lingual layer based on universal PoS frequencies, and a mono-lingual layer based on PoS histograms was added to the graph. The results were compared to the IRL and TSVM methods with identical cross-lingual features (text statistics and universal PoS tags) and the cross-lingual SCL baseline. Figures 7.7, 7.8, and 7.9 show this comparison for the two, four, and nine genre tasks respectively.

It is obvious from the results that the label propagation strongly benefits from the additional PoS-based features and resulting graph layers. In fact, with decreasing source input, the algorithm achieves better accuracy for 17, and better F1-Scores for 16 of the 18 tasks, when compared to the results in Figures 7.4, 7.5, and 7.6. The

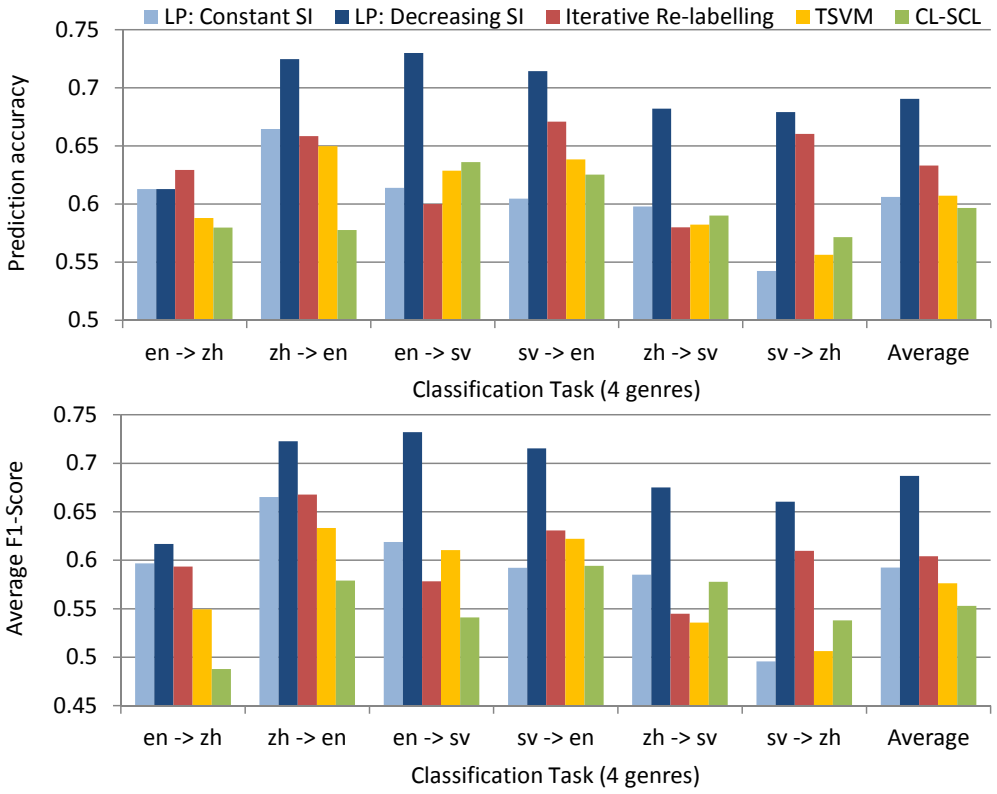

Figure 7.8: Same as Figure 7.7, but for the four genre tasks.

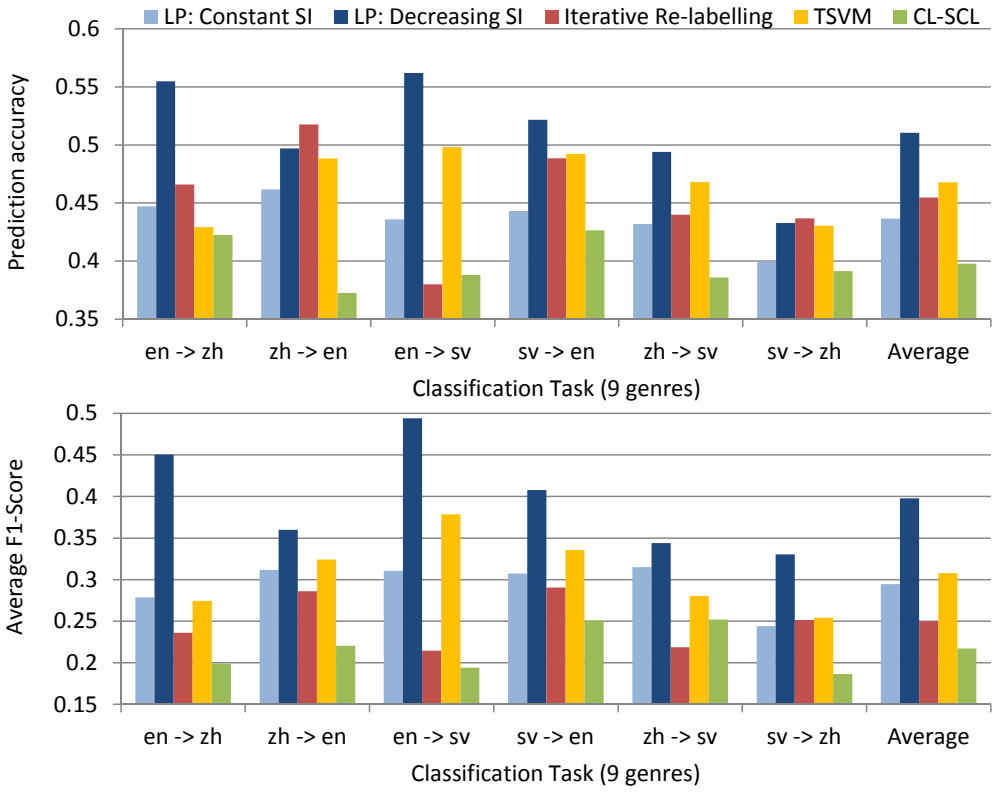

Figure 7.9: Same as Figure 7.7, but for the nine genre tasks.

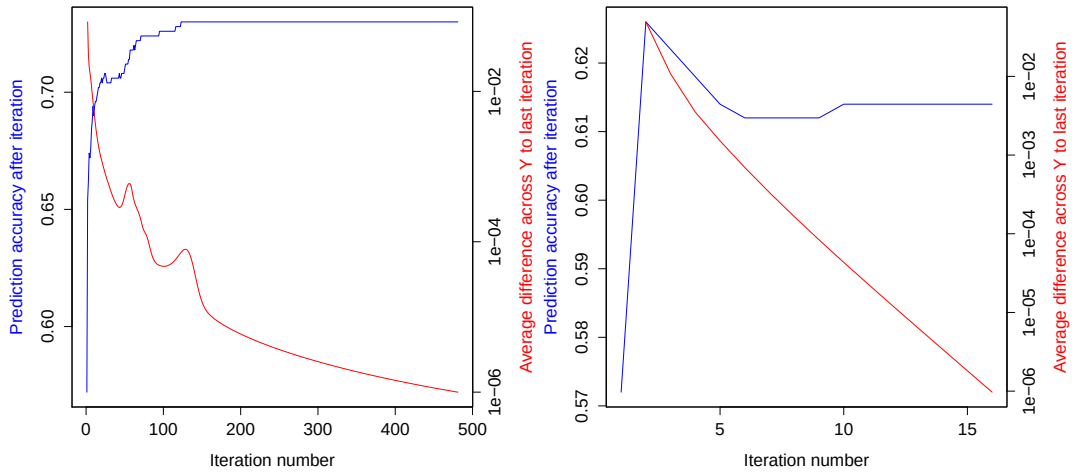

Figure 7.10: For each iteration in the four genre en→sv task, the graphs show prediction accuracies (blue) and average differences in Y (red) to the previous iteration. The graphs on the left and right show results for the variants with decreasing and constant source input, respectively.

improvements are particularly strong for the fine-grained task, as the average accuracy over all nine genre tasks increased from 38.1% to 51.0%. The benefits are similar, if less strong, when source input is kept constant. When compared to the other classifiers, label propagation with decreasing source input yields very competitive results. For two genres, both accuracies and F1-Scores are almost identical to those of the IRL method and mostly better than those of the other baselines. For the more fine-grained tasks (four and nine genres), the algorithm outperforms the other methods for most source-target combinations, often strongly. This is particularly true for the average F1-Score metric. It is also the only method that achieves accuracies beyond 70% with four genres and beyond 55% with nine genres. The variant with stable source input cannot achieve such strong results, but performs well when compared to other methods on the four and nine genre tasks.

Another evaluation criteria is the number of iterations it takes for the algorithm to converge. Figures 7.10, 7.11, and 7.12 show the development of prediction accuracies and average differences in Y to the previous iteration throughout the iterative processes. The latter is used as the convergence criteria. The results reveal a strong difference in convergence speed between the two label propagation variants. The algorithm required about 20 iterations to reach the threshold if source input was kept constant. This increased to several hundred iterations with decreasing source input. The reason is that

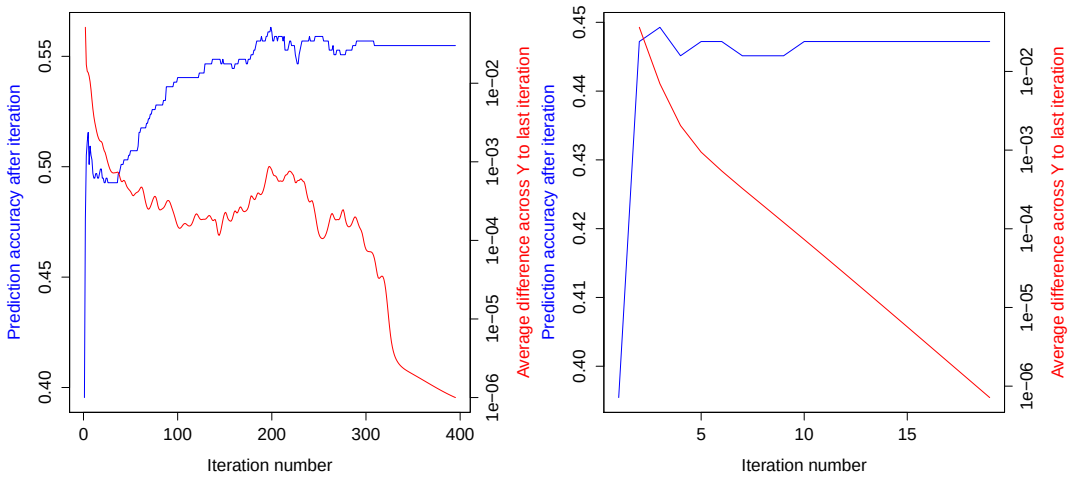

Figure 7.11: Same as Figure 7.10, but for the nine genre en→zh task.

the input from the genre-annotated source language nodes, whose label probabilities are constant, has a regulating effect. This leads to quick convergence, which is desirable, but can often prevent a better solution. Note that while the algorithm converged in all my experiments, there is no formal proof that this is always the case for any of the extended versions of label propagation described in this chapter.

Figures 7.10, 7.11, and 7.12 also illustrate the improvement in accuracy that is gained from unlabelled texts in the target language. The relatively low starting points of the blue lines correspond to the performances after the first iteration, where only source-target edges are used. For all tasks, a large improvement can be observed after the second iteration. This is due to a combination of the effect of target-target edges and the adjustments made through genre-specific layer weights (see Section 7.2.4 for a more detailed evaluation). While for most tasks, the final accuracy was close to the maximum observed throughout the iterative process, this was not always the case. Figure 7.12 shows that results in the early iterations are better than that after convergence, in particular with decreasing source input. A stronger regulation may help in such cases, but inhibit possible positive developments, as can be observed in Figure 7.11.

7.2.4 Detailed Results

The results in Section 7.2.3 demonstrate that the label propagation method as described in this chapter performs very well. In order to determine which of the variants of the algorithm actually yield benefits for a CLGC task, experiments were designed to

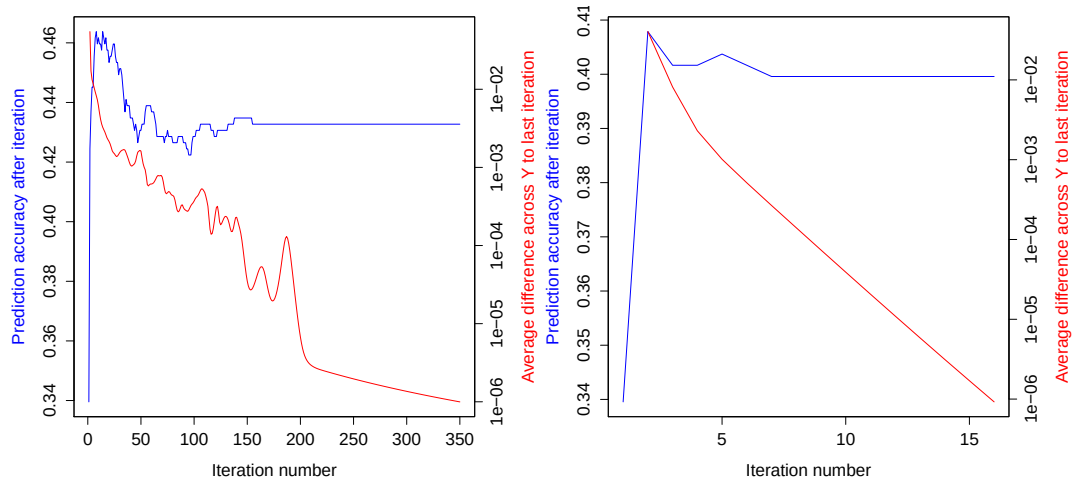

Figure 7.12: Same as Figure 7.10, but for the nine genre sv→zh task.

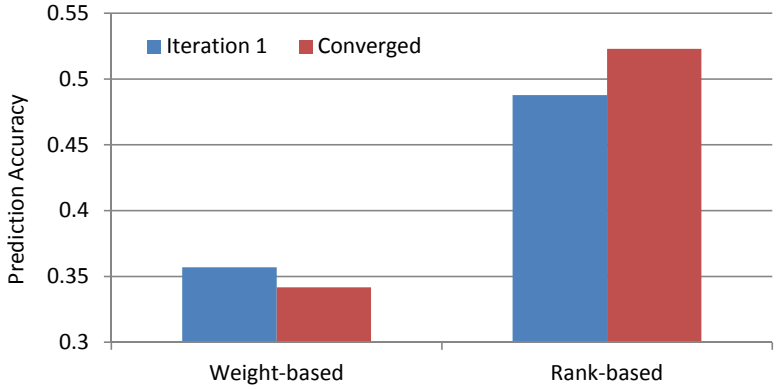

Figure 7.13: Prediction accuracies of label propagation classifiers with distance-based (left) and rank-based (right) edge weights (6 genre classes).

evaluate one extension at a time, keeping everything else constant. For this task, the constant source input variant was used. Firstly, the impact of rank-based weights was evaluated. To this end, the label propagation method was implemented as described in Section 7.1, that is with multiple layers, learned layer weights, and confidence values. In the first run, the initial layer weights were based on Euclidean distances, as explained in Section 7.1.1. In the second run, rank-based distances (Section 7.1.3) were used.

Figure 7.13 shows that distance-based edge weights produce a relatively poor classification result. As the accuracy after the first iteration (where labels are propagated from source language texts to target language texts only) is low, the subsequent cannot improve the outcome. In comparison, the classifier using rank-based edge weights achieves a significantly better accuracy. Furthermore, the initial labelling of the target

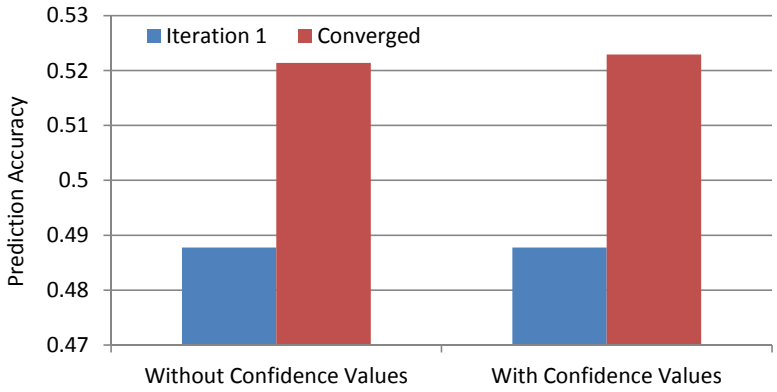

Figure 7.14: Prediction accuracies of label propagation classifiers with (right) and without (left) confidence values factoring into the impact of a node.

language texts is good enough for a subsequent significant improvement by the iterative algorithm.

Next, it was evaluated whether or not the label propagation algorithm benefits from computing confidence values to strengthen the impact of nodes with unambiguous genre predictions (see Section 7.1.4). The baseline for this method was identical to the described algorithm, but confidence values were ignored (i.e., $c_j = 1$ for all nodes j).

The results are illustrated in Figure 7.14. Surprisingly, there is no statistically significant difference between the two. This is unlike the iterative re-labelling algorithm, where choosing only texts with high label confidence for re-training boosted the classifier performance. A possible reason for this is the comparatively soft way that low label confidence is punished. While the iterative re-labelling algorithm excludes texts that fall below a threshold completely, the label propagation method merely reduces their impact compared to texts with more confidence in their labels.

One of the most central adjustments of the original algorithm proposed by Zhu and Ghahramani (2002) is the introduction of several graph layers. Two of the benefits are easy adjustments based on the available resources for a language pair and genre-specific feature set weighing, which allows for interpretation of the (cross-lingual) characteristics of a genre. Experiments were carried out to determine whether this translates to an increased classification accuracy as well. A baseline was implemented with only one graph layer. This uses the same cross-lingual and mono-lingual features as the multi-layer variant, but they are combined into a cross-lingual and a mono-lingual feature set. The edge weights are then computed using the distance ranks in this combined feature space. Source language nodes are connected to target language nodes using the

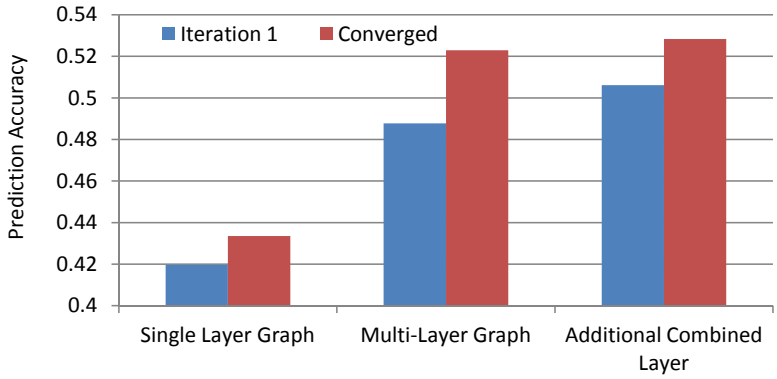

Figure 7.15: Prediction accuracies of three variants of the label propagation classifier.

cross-lingual feature set, while connections between target language nodes exploit the mono-lingual set. Therefore, at most one directed edge points from any given node j to another node i.

Figure 7.15 shows that the classifier can benefit strongly from separate graph layers, although it makes use of the same features as the single-layer version. However, looking at how exactly the accuracies were achieved, there are strong differences between the two, and the multi-layer graph does not outperform the single-layer graph for every genre class. Figure 7.16 shows the number of correctly predicted texts for both the single-layer and the multi-layer graphs (i.e. combined vs. separated feature sets). It is clear that the recall values for some genres benefit from the separated feature sets, while the opposite is the case for other genres. The difference is clearest for biographies and letters. While the former are predicted correctly more easily with a combined feature set, the latter are clearly harmed by this.

This observation motivated another experiment, where the combined cross-lingual and mono-lingual feature sets were added as another layer to the label propagation graph, in addition to the layers derived from the separated feature sets. The resulting accuracy is also illustrated in Figure 7.15. The additional layer yielded a slightly higher accuracy than that achieved by a multi-layer graph without the combined feature set. While the difference is not statistically significant, it is interesting to see how it breaks down to correctly predicted texts by genre.

As Figure 7.17 illustrates, the positive difference is entirely due to a higher recall in the biography class. On the other hand, there is almost no change in the number of correctly classified texts in the other genres. This is surprising in particular for letters, considering the large difference between single-layer and multi-layer graphs for this

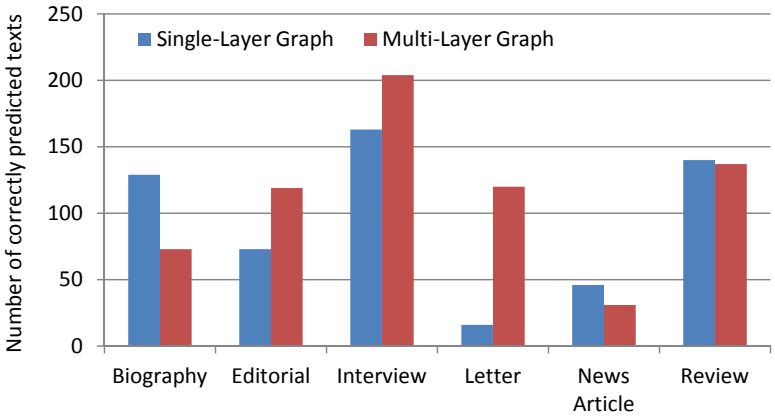

Figure 7.16: Number of correctly predicted texts in each genre class achieved by a single-layer and a multi-layer classifier.

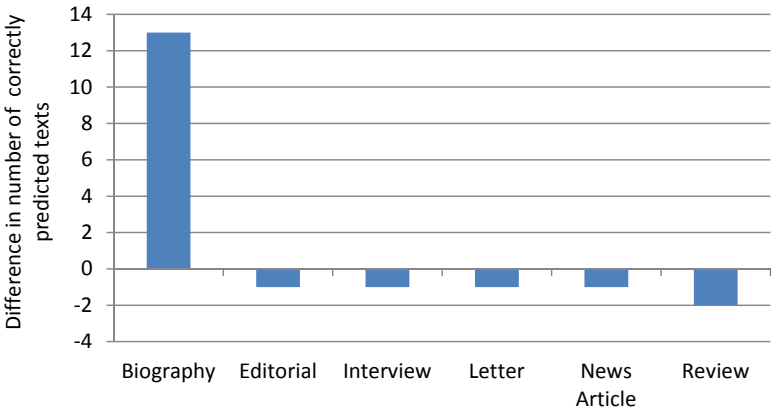

Figure 7.17: Difference in the numbers of correctly classified texts by a multi-layer label propagation classifier after adding an additional combined feature layer.

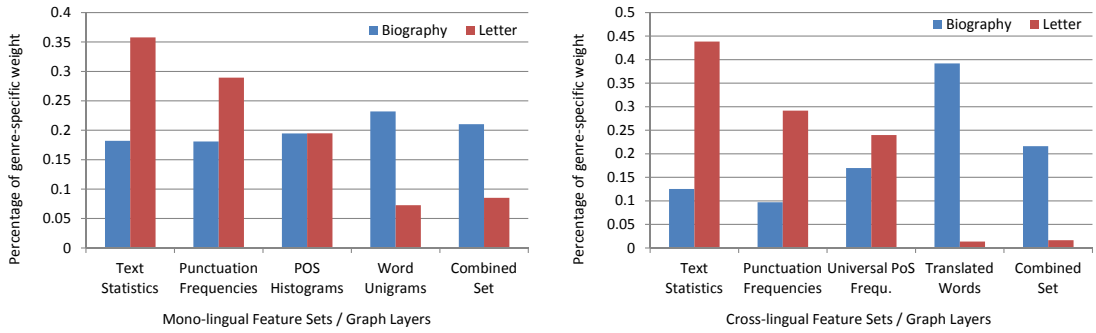

Figure 7.18: Genre-specific cross-lingual (left) and mono-lingual (right) layer weights for the biography and letter classes.

genre shown in Figure 7.16. However, looking at the learned layer weights provides an explanation.

Figure 7.18 shows the genre-specific layer weights that were learned by the algorithm for biographies and letters. These are taken from the modification matrix M (described in Section 7.1.2) after convergence. They can be interpreted as a measure of how coherent texts within a genre are with respect to different feature spaces. While biographies make considerable use of the added combined feature graph layer, letters almost completely ignore it, especially for source-target edges. This means that nodes with a high probability of belonging to the letter class will mostly propagate their labels through the other layers. This explains the beneficial effect of the added layer on the number of correctly predicted biographies and the small impact on the letter class.

More evidence for the positive effect of added layers can be found in Figure 7.19. This shows the classification accuracies for different assumptions of available resources in the target language. The most basic approach uses only text statistics to bridge the language gap. The others assume the availability of PoS taggers, cross-lingual punctuation mappings, and/or machine translation. All of the experiments shown in Figure 7.19 exploit text statistics, punctuation frequencies, and word unigram features within the target language. Where universal PoS tags are used across languages, PoS histograms are exploited to compute edge weights between target language nodes. One exception is the MT-only method shown on the left: This is a simple, though resource-intensive, baseline using translated words and target language word unigrams only.

In general, adding additional layers improved the classification accuracy in these experiments. This shows that the algorithm handles the added features well. It also

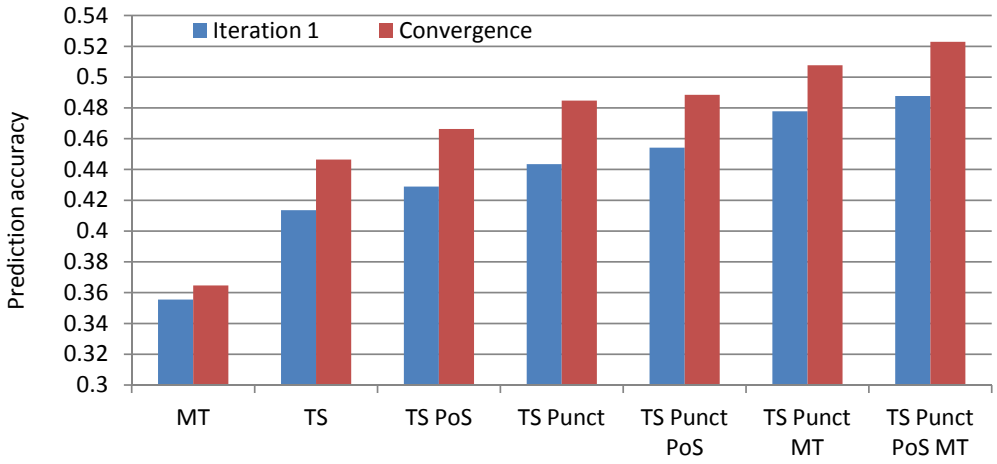

Figure 7.19: Prediction accuracies of the label propagation algorithm with different feature sets. **MT**: Machine translation based features. **TS**: Text statistics. **Punct**: Punctuation features. **PoS**: Part of Speech features.

indicates that genres can indeed be predicted with a variety of different types of features. Finally, it demonstrates that even where few resources in the target language are available, it is possible to achieve an accuracy across languages, which is good enough to allow for further improvement through unlabelled target language texts.

As mentioned in Section 7.1.6, the label propagation algorithm is transductive in nature. Figure 7.20 shows the performance of an SVM model trained on a subset of the target language texts, after they received labels through the graph. The features used for this second step are the PoS histograms proposed by Feldman et al. (2009) for mono-lingual genre classification. Unsurprisingly, the accuracy is lower than that of the transductive algorithm. However, the difference is relatively small. To put this result into perspective, Figure 7.20 also shows the accuracy achieved by an SVM model directly trained on source language texts, using the universal PoS frequencies as features. The difference highlights the advantage of exploiting target language texts for training where available, even if an inductive classification model is required.

As a proof-of-concept, the classifier was also tested with data from the CIIL corpus (see 2.4), to further evaluate whether the method works for poorly-resourced target languages. Here, English texts from the British National Corpus were used as source nodes while Tamil and Malayalam texts were used as target nodes. Note that the genre distributions in the BNC and the CIIL corpora are very different. Therefore, in addition to the usual scenario, a second option was evaluated. This requires the input from an oracle, which was also proposed by Zhu and Ghahramani (2002). This oracle reveals

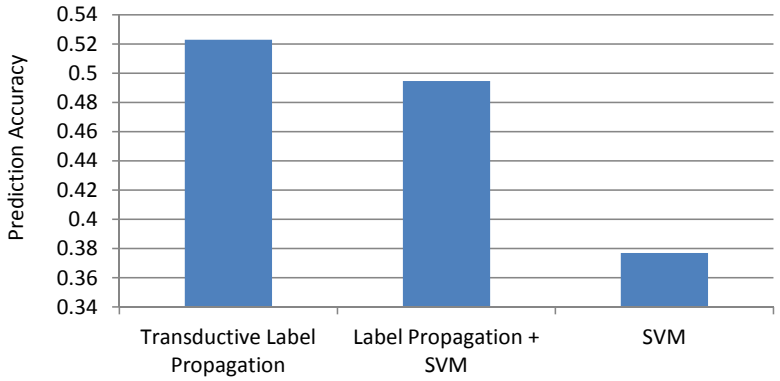

Figure 7.20: Prediction accuracies of the transductive label propagation algorithm (left) and an SVM trained on its output in the target language (middle). The right bar shows an SVM directly trained on the source language.

the genre distribution in the target language. This is used to scale the label probabilities of target language nodes after the iterative process. Note that the oracle does not need to know the genre label of any target language text.

As the target language is relatively poorly-resourced, the classifier assumes minimal knowledge of the target language. Therefore, only a reduced set of text statistics without paragraph-based features was used to compute the cross-lingual edge weights. Figures 7.21 and 7.22 show the results, compared to those of random guess classifiers, with and without use of the oracle. In both cases and for both target languages, the label propagation method outperforms the respective baseline, which provides further evidence for the thesis of this project. It is also evident that the input from the oracle boosted the classification performance. While such an oracle is typically unavailable in practice, it shows that a correct intuition of the genre distribution in the target language can be exploited by the label propagation method to further improve results.

7.3 Discussion

To summarize the experimental results in Sections 7.2.3 and 7.2.4, it can be observed that label propagation using a multi-layered graph is a suitable semi-supervised classifier for cross-lingual genre classification problems, in particular where cross-lingual resources are available. In such cases, it can outperform the same algorithm using a single-layered graph. Rank-based edge weights both allow complexity optimization and improved results in my experiments. The use of prediction confidence values, on the other hand,

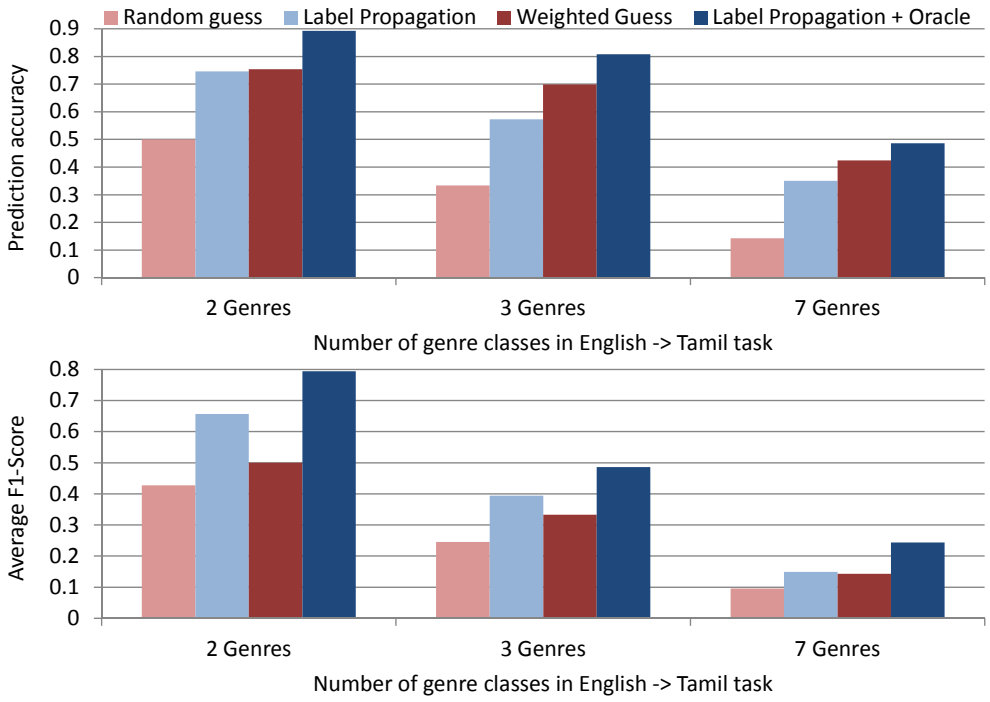

Figure 7.21: Prediction accuracies (top) and average F1-Scores (bottom) for the English to Tamil classification task with two, three, and seven genre classes. The light coloured bars correspond to classifiers without knowledge of the target language genre distribution. The dark coloured bars show results of classifiers that rely on an oracle to reveal them.

did not prove to help the accuracy, which is different from what was observed for the iterative re-labelling method.

The experiments carried out with the label propagation algorithm provide further evidence that genres correlate with simple text features which are comparable across languages. This can be observed for all of the various genre palettes and language pairs examined. It was also demonstrated, however, that added linguistic resources can improve results. The algorithm proposed here allows adding feature sets based on what resources are available for a language pair and the experimental results show that it handles the additional knowledge well by assigning genre-specific weights to graph layers.

These weights also reveal characteristics of the different genres. As can be seen in Figure 7.18, letters and biographies are close to texts of the same genre in different feature spaces. While letters rely on text statistics, universal PoS tags, and punctuation features to communicate their labels across languages, biographies use mostly the edges derived from machine translation. It can also be observed that the impact of the same

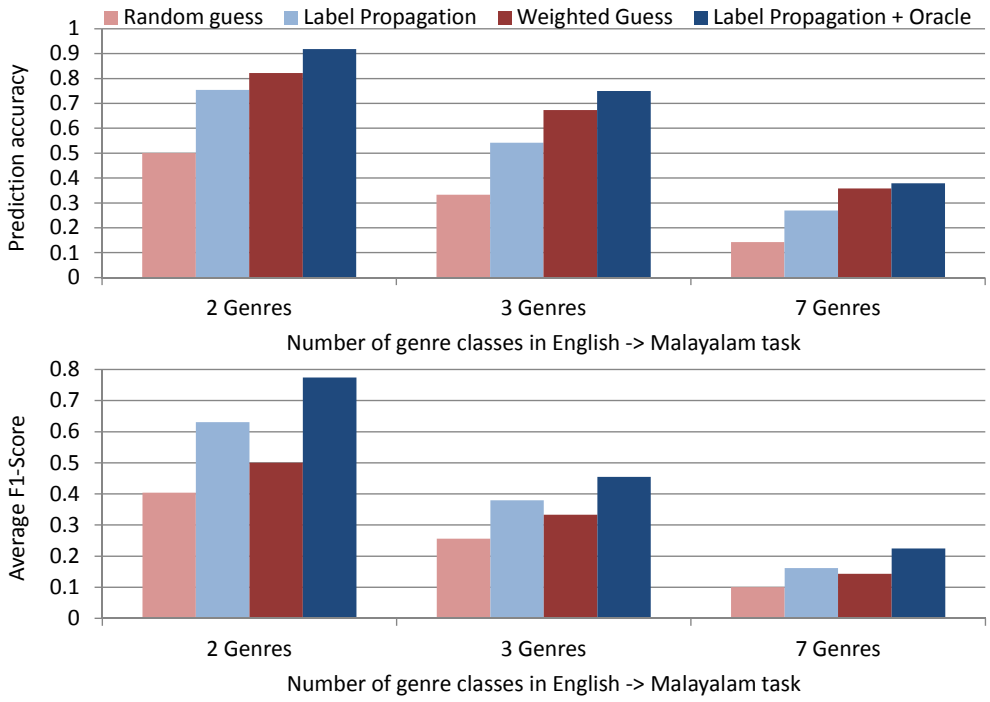

Figure 7.22: Same as Figure 7.21, but with Malayalam as the target language.

features can differ for mono-lingual and cross-lingual layers (cf. text statistics and punctuation features for biographies).

It is important to note that these weights are an interpretation of the algorithm and may not reflect the true characteristics of a genre with respect to the different feature sets, unless all target language instances are labelled correctly with 100% confidence. Nevertheless, they provide some evidence that one genre can be well described along dimensions that may be irrelevant to other genres. Furthermore, the difference between the cross-lingual and the mono-lingual weight distribution indicates that features that describe a genre well within a language, do not necessarily make for good genre predictors across languages.

This also means that classification algorithms, which perform a supervised selection or weighing of features based on the labels in the training set should be used with care for this task. This is because features that work well to predict a text's genre in the source language, might be harmful cross-lingually, since the target language is not considered during training. The approach taken here is therefore to balance the weight of feature sets initially and learn weights based on the label predictions in the target language.

Chapter 8

Conclusion

This dissertation presents the first work on cross-lingual genre classification (CLGC). I showed that simply extractable text features exist that correlate with genre similarly in different languages, and which can be exploited as cross-lingual features. I also demonstrated that it is possible to achieve good results in CLGC tasks without resource-intensive machine translation techniques or supervised tools in the target language. That is not to say that such resources, where available, cannot improve classification results. In fact, some of the experimental results show that additional cross-lingual knowledge and/or knowledge about the target language can be incorporated to enhance CLGC methods. However, they are not required, which makes the methods presented here suitable for poorly-resourced target languages.

As this is the first work on CLGC, suitable data for analyses and experiments had to be identified. For this project, ten publicly available text corpora were used and experiments were carried out with eleven languages: Chinese, Danish, English, French, German, Italian, Malayalam, Portuguese, Spanish, Swedish, and Tamil. While the source corpora were not designed for the task of CLGC, their meta-data allows inference about the genre of the included texts. As the data is publicly available, and in some cases free of charge, the corpora suggestion and pre-processing steps introduced in this dissertation are hoped to encourage and facilitate further research in the area of CLGC. While more suitable text collections will hopefully be available in the future (see Chapter 9), the proposed data sets can be considered a starting point for researchers to develop and compare methods.

Like prior work on cross-lingual techniques, some of the approaches presented here are semi-supervised classifiers and exploit both labelled texts in the source language and unlabelled texts in the target language. I provided evidence of this yielding better

results when compared to a strictly supervised approach exploiting source language texts only. This is particularly true when text features are separated into two groups: (1) Cross-lingual features that help to bridge the language gap, and (2) those that correlate with genre in the target language and can be used to refine and improve the initial cross-lingual prediction. While the latter set can include cross-lingual features also, it is less restricted, as these features have to work only within the target language.

The features used for the experiments in this project fall into four broad sets with different levels of resource requirements: Text statistics, punctuation features, PoS-based features, and word frequencies. The first two sets are used both across languages and within the target language. The latter two are split into task-specific versions, with the cross-lingual feature sets exploiting a universal PoS tag mapping and machine translation, respectively. The experimental results presented here show that all four types of features can improve classification results if they can be combined in a suitable way by the cross-lingual classifier. While many of the features I used were suggested for mono-lingual genre classification before, some are new for this task. A particularly interesting finding is the beneficial effect of PoS tag frequencies for CLGC tasks. While PoS tags have previously been shown to be helpful in distinguishing between genre categories, grammar differences between languages make a direct comparison difficult. Thanks to recent efforts in cross-lingual PoS tag mappings, such features can now be utilized for a wide range of language pairs.

Furthermore, I presented different algorithms to make use of such features. One approach exploits cross-lingual features to bridge the language gap and subsequently applies iterative target language adaptation with a target language specific feature set in order to improve accuracy. This iterative re-labelling method also employs a selection of target language texts with high confidence in their predicted labels. These are then used to train the classifier for the next iteration of the algorithm. For most of the experiments with broad genre categories (two or four genres), the approach performed equally well or better than full text translation combined with mono-lingual classification, while requiring less resources. It also outperformed the other baselines for most tasks in this scenario. However, I also showed that it performs considerably worse when a more fine-grained genre palette is used. This is especially true when considering the average F1-Score metric, which shows that the method over-predicts dominant genre classes. Furthermore, I demonstrated that this relatively simple approach cannot benefit significantly from additional cross-lingual features. It can therefore be considered to be more suited for classifying texts into broad genre categories for poorly-resourced

language pairs, which allow only a restricted set of cross-lingual features.

Arguably, one of the reasons why the iterative re-labelling method cannot benefit from addition cross-lingual resources is that it cannot weigh or select features based on their cross-lingual correlation with genre. This is because no labels are assumed to be available for the target language. However, I showed that such a weighing or selection can be achieved if labels are available for texts from more than one language, even if these do not include the target language. One approach is to train a supervised classifier on a set of texts from different languages. I also presented a cross-lingual feature selection method, which can find a lower dimensional representation for the source and target language texts independent of either language. In the experiments with eight European languages, both methods outperform the respective baselines of using a single source language or all available cross-lingual features. This shows the importance of evaluating the cross-lingual predictive power of input variables with respect to genre. A good predictor for genre in one source language is not necessarily a good cross-lingual predictor. I demonstrated that, for European languages at least, comparable corpora can be used to automatically identify predictive features from a set of candidates.

However, genre-annotated texts from multiple languages may not be available. Moreover, the cross-lingual feature selection does not take the target language into consideration. This may be a problem if there are differences between source and target languages with respect to genre. To address this, I propose an alternative graph-based classifier based on the label propagation algorithm. For each feature set, a separate layer in a graph is created, where each text is represented by a node. These nodes then propagate their genre label beliefs through the edges on different layers. During this iterative process, the algorithm learns which feature sets connect texts of a given genre well and increases the weights of the corresponding edges. This means that each text propagates its label mostly through the edges based on feature sets that are believed to be predictive for its known or predicted genre class. Both source and target languages are taken into consideration here. While no manually annotated texts can be employed for the latter, the algorithm estimates the usefulness of a feature set based on the predicted labels from previous iterations.

I demonstrated that the multi-layered graph based classifier can benefit from additional resources, if they become available. The experimental results show increased accuracies and F1-Scores if new feature sets are added. Furthermore, the method performs comparatively well for fine-grained classification tasks with several genre

categories. These two findings are unlike the observations made for the iterative relabelling method. This suggests that the label propagation algorithm is the better choice for well-resourced language pairs and/or large target genre palettes. On the other hand, the approach is inherently transductive, which means that target language texts have to be available at training time in order to be classified. I proposed and evaluated a remedy by training an inductive classifier on a subset of target language texts and their newly predicted genre labels. However, for online problems, a more direct solution may be more appropriate and yield better results.

In conclusion, this project explored different data sets, features, and algorithms for the task of cross-lingual genre classification. It was shown that good results can be achieved with no or few cross-lingual resources, by exploiting labelled texts in the source language and unlabelled texts in the target language. Beyond providing working methods for practical applications, these encouraging findings are hoped to inspire future work in the field.

Chapter 9

Future Work

There are several possible directions for future work. They can be grouped into the three broad fields of data, features, and algorithms.

One of the most crucial areas for the future of CLGC may be the construction of a publicly available corpus with genre-annotated texts in several languages, which goes beyond the data sets used for this project. Unfortunately, the term genre still carries a lot of ambiguity, even within a single language. Therefore, creating such a corpus would require a collaborative international effort of genre researchers in order to reach a consensus on the text sources, genre palette, sampling methodology, annotation rules, and formatting. Such a project would certainly be challenging, as it may involve obtaining copyright licences in different countries. However, the resulting dataset would allow researchers to create and evaluate new CLGC methods, and enable a straightforward and meaningful comparison of algorithms. While in this project a first step was made by locating existing publicly available data that can be used for this task, a multi-lingual corpus created specifically for genre classification could overcome the restrictions of the text collections used here. For instance, more languages could be represented, the genre palette could be created based on a given application (e.g. information retrieval or text summarization), and texts could be annotated with other potentially useful meta-data.

Another area that would benefit from additional work is the identification of further cross-lingual genre-revealing features. The features presented in Chapter 3 focus on the texts themselves, which may be all that is available to a classifier. However, prior work on mono-lingual genre classification has involved several other types of features, which may be suited to bridge the language gap (see Section 1.2). An example are image features, which can distinguish texts by their layout. Certain genre-based layout

conventions may hold across different languages, thus experiments are needed to show whether image features can benefit CLGC. In web genre classification, markup features have been used to model the structure of a document. Since HTML or CSS tags are language-independent, they can be directly compared across languages. However, this is not to say that the same web genres in different languages use markup identically. Therefore, further research would help to find language-independent and language-specific patterns.

Even as far as text-based features are concerned, more ground is yet to be covered. In this project, I focussed on low-level features, which require little or no resources. While PoS tags were explored and evaluated as cross-lingual predictors, more linguistically sophisticated features may help to improve classification performances. To this end, parsers could be utilized to explore genre-specific, syntactic similarities between languages. While such high-level features are unlikely to work across languages with strong grammatical differences, they might be beneficial when working with closely related language pairs.

Furthermore, experiments with feature sets designed to correlate with genre facets would be very interesting. The multi-layered label propagation algorithm presented in Chapter 7 has been shown to assign genre-specific layer weights, both for cross-lingual and target language specific feature sets. Where features are known to be predictive of certain facets, they can be grouped and used to assign edge weights of separate graph layers. Such knowledge can come from prior work on text characteristics such as target audience, reading level, register, sentiment, topic, objectivity, and others. Where the respective features can be extracted from both source and target languages, a classifier might learn which facets are found in genres of either language. If they are specific to the target language (e.g. word frequencies) they might be used to improve classification accuracy after the language gap has been bridged. The goal would be an algorithm that learns which facets are relevant for which genre and uses this knowledge to classify texts more robustly.

Finally, additional effort might be required to find classification algorithms suitable for the task of CLGC. The methods presented in this dissertation can serve as a starting point for further developments, and/or as a baseline for new approaches. New algorithms could either aim to be general solutions or more specialized approaches to address certain sub-problems of CLGC. They could, for example, concentrate on a specific language pair, or a specific target language (regardless of the source language). They could also address the potential issue of very different genre distributions in the source

language set and the target language set. This problem was mentioned in Section 3.5, and a possible solution was proposed there for finding suitable feature scaling values in such cases. However, this is yet to be evaluated empirically, and difference in genre distributions bring further problems, which the methods of this dissertation do not address specifically.

A few pointers were already provided for further developments of the algorithms presented in this dissertation. For example, the cross-lingual feature selection method discussed in Chapter 6 might benefit from an improved threshold search strategy and/or a set-based selection rather than feature ranking. The label propagation algorithm (Chapter 7) might be adapted to enable the classification of new texts more naturally, preferably retaining the benefit of separate feature sets for induction. Another interesting extension could be the inclusion of further unlabelled texts written in the source language, or even a third, non-target language for which no labelled data is available. While such data can already be represented as nodes in the graph, adaptations are likely required to optimally exploit these additional resources. Finally, all methods might be adjusted in order to cope better with imbalanced class distributions, beyond what was already described. Solutions could be derived from a large body of research into this sub-field of machine learning (e.g., Galar et al. 2012; Wang and Yao 2012; Lin and Chen 2013).

Bibliography

Argamon, S., Koppel, M., and Avneri, G. (1998). Routing documents according to style. In *Proceedings of the First International Workshop on Innovative Information Systems*, pages 85–92. Citeseer.

Bagdanov, A. D. and Worring, M. (2001). Fine-grained document genre classification using first order random graphs. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition, ICDAR2001*, pages 79–83. IEEE.

Bel, N., Koster, C., and Villegas, M. (2003). Cross-lingual text categorization. In Koch, T. and Slvberg, I., editors, *Research and Advanced Technology for Digital Libraries*, volume 2769 of *Lecture Notes in Computer Science*, pages 126–139. Springer Berlin / Heidelberg.

Berninger, V., Kim, Y., and Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. In *Proceedings of the 2008 BCS-IRSG conference on Corpus Profiling*, pages 2–2. British Computer Society.

Biber, D. (1991). *Variation across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation*. Cambridge University Press.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128. Association for Computational Linguistics.

Boitet, C., Blanchon, H., Seligman, M., and Bellynck, V. (2010). MT on and for the web. In *Proceedings of the 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–10. IEEE.

Braslavski, P. (2004). Document style recognition using shallow statistical analysis. In *Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, pages 1–9.

Bruce, I. (2011). Evolving genres in online domains: The hybrid genre of the participatory news article. In *Genres on the Web*, pages 323–348. Springer.

Burnard, L. (2000). Reference guide for the British National Corpus (World Edition). http://www.natcorp.ox.ac.uk/.

Central Institute of Indian Languages (2011). CIIL Corpus. http://ltrc.iiit.ac.in/corpus/corpus.html.

Chaker, J. and Habib, O. (2007). Genre categorization of web pages. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW '07, pages 455–464. IEEE Computer Society.

Chang, C. and Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cronen-Townsend, S. and Croft, W. B. (2002). Quantifying query ambiguity. In *Proceedings of the second international conference on Human Language Technology Research*, pages 104–109. Morgan Kaufmann Publishers Inc.

Crowston, K., Kwaśnik, B., and Rubleske, J. (2011). Problems in the use-centered development of a taxonomy of web genres. In *Genres on the Web*, pages 69–84. Springer.

De Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Fayyad, U. and Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. page 10221027.

Feldman, S., Marin, M. A., Ostendorf, M., and Gupta, M. R. (2009). Part-of-speech histograms for genre classification of text. In *Proceedings of the 2009 IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing, pages 4781–4784. IEEE Computer Society.

Ferizis, G. and Bailey, P. (2006). Towards practical genre classification of web documents. In *Proceedings of the 15th international conference on WWW*, pages 1013–1014. ACM.

Finn, A. and Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518.

Francis, W. N. and Kucera, H. (1979). Brown Corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Freund, L., Clarke, C. L. A., and Toms, E. G. (2006). Towards genre classification for IR in the workplace. In *Proceedings of the 1st international conference on Information interaction in context*, pages 30–36. ACM.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4):463–484.

Gliozzo, A. and Strapparava, C. (2006). Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 553–560. Association for Computational Linguistics.

Goldstein, J., Ciany, G. M., and Carbonell, J. G. (2007). Genre identification and goal-focused summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 889–892. ACM.

Gustafson-Capková, S. and Hartmann, B. (2006). Manual of the Stockholm Umeå Corpus version 2.0. Technical report, Dept. of Linguistics, Stockholm University.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Herdan, G. (1960). *Type-token mathematics*, volume 4. Mouton.

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142. Springer-Verlag.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of The Sixteenth International Conference on Machine Learning*, volume 99, pages 200–209.

Kanaris, I. and Stamatatos, E. (2007). Webpage genre identification using variable-length character n-grams. In *Proceedings of the 19th IEEE International Conference on Tools with AI*, pages 3–10.

Karlgren, J. (2011). Conventions and mutual expectations. In *Genres on the Web*, pages 33–46. Springer.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075. Association for Computational Linguistics.

Kessler, B., Nunberg, G., and Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics.

Kim, Y. and Ross, S. (2008). Examining variations of prominent features in genre classification. In *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, HICSS '08, pages 132–. IEEE Computer Society.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32:485–525.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Lee, Y.-B. and Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM*

SIGIR conference on Research and development in information retrieval, SIGIR '02, pages 145–150. ACM.

Lim, C. S., Lee, K. J., and Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Inf. Process. Manage.*, 41(5):1263–1276.

Lin, W.-J. and Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics*, 14(1):13–26.

McEnery, A. and Xiao, Z. (2004). The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In *LREC*. European Language Resources Association.

Mehler, A., Sharoff, S., and Santini, M. (2010). *Genres on the Web: Computational Models and Empirical Studies*, volume 42. Springer.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2012). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-1.

Meyer zu Eissen, S. and Stein, B. (2004). Genre classification of web pages. In Biundo, S., Frhwirth, T., and Palm, G., editors, *KI 2004: Advances in Artificial Intelligence*, volume 3238 of *Lecture Notes in Computer Science*, pages 256–269. Springer Berlin / Heidelberg.

Müller, F. H. (2004). Stylebook for the Tübingen partially parsed corpus of written German (TüPP-D/Z). In *Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen*, volume 28, page 2006.

Oberlander, J. and Nowson, S. (2006). Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 627–634. Association for Computational Linguistics.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 751–760. ACM.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on*

Empirical methods in natural language processing - Volume 10, EMNLP '02, pages 79–86. Association for Computational Linguistics.

Petrenz, P. (2009). Assessing approaches to genre classification. MSc thesis, School of Informatics, University of Edinburgh.

Petrenz, P. (2012). Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–21. Association for Computational Linguistics.

Petrenz, P. and Webber, B. (2011). Stable classification of text genres. *Computational Linguistics*, 37:385–393.

Petrenz, P. and Webber, B. (2012a). Label propagation for fine-grained cross-lingual genre classification. In *Proceedings of the NIPS Workshop on Cross-Lingual Technologies (xLiTe)*.

Petrenz, P. and Webber, B. (2012b). Robust cross-lingual genre classification through comparable corpora. In *The 5th Workshop on Building and Using Comparable Corpora*, page 1.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Pivovarova, L., Huttunen, S., and Yangarber, R. (2013). Event representation across genre. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 29–37. Association for Computational Linguistics.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Prettenhofer, P. and Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1118–1127. Association for Computational Linguistics.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008). Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the 6th international conference on language resources and evaluation (LREC 2008)*.

Rigutini, L., Maggini, M., and Liu, B. (2005). An EM based training algorithm for cross-language text categorization. In *Proceedings of the Web Intelligence Conference*, pages 529–535.

Romanski, P. (2013). *FSelector: Selecting attributes*. R package version 0.19.

Rose, T., Stevenson, M., and Whitehead, M. (2002). The Reuters corpus volume 1 – from yesterday's news to tomorrow's language resources. In Shavlik, J. W., editor, *Proceedings of the Third Internaional Conference on Language Resources and Evaluation*.

Rosso, M. A. and Haas, S. W. (2011). Identification of web genres by user warrant. In *Genres on the Web*, pages 47–67. Springer.

Sandhaus, E. (2008). New York Times corpus: Corpus overview. LDC catalogue entry LDC2008T19.

Santini, M., Mehler, A., and Sharoff, S. (2011). Riding the rough waves of genre on the web. In *Genres on the Web*, pages 3–30. Springer.

Santini, M., Power, R., and Evans, R. (2006). Implementing a characterization of genre for automatic genre identification of web pages. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 699–706. Association for Computational Linguistics.

Scholl, P., Domínguez García, R., Böhnstedt, D., Rensing, C., and Steinmetz, R. (2009). Towards language-independent web genre detection. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1157–1158. ACM.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

Sharoff, S. (2007). Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*.

Sharoff, S. (2010). In the garden and in the jungle: Comparing genres in the BNC and internet. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer Berlin / New York.

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web Library of Babel: Evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 3063–3070. European Language Resources Association (ELRA).

Sindhwani, V. and Keerthi, S. S. (2006). Large scale semi-supervised linear SVMs. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 477–484. ACM.

Snyman, D. P., van Huyssteen, G. B., and Daelemans, W. (2012). Cross-lingual genre classification for closely related languages. In *Proceedings of the Twenty-Second Annual Symposium of the Pattern Recognition Association of South Africa*, pages 133–137.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000a). Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, pages 808–814. Association for Computational Linguistics.

Stamatatos, E., Kokkinakis, G., and Fakotakis, N. (2000b). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Stein, B., zu Eissen, S. M., and Lipka, N. (2011). Web genre analysis: Use cases, retrieval models, and implementation issues. In *Genres on the Web*, pages 167–189. Springer.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.

Stubbe, A. (2006). Klassifikation von texten nach genre. Dissertation (German), LMU Munich.

Swales, J. M. (1990). *Genre Analysis: English in academic and research settings*. The Cambridge applied linguistics series. Cambridge University Press.

Thomson, E. A., White, P. R., and Kitley, P. (2008). Objectivity and hard news reporting across cultures. *Journalism Studies*, 9(2):212–228.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the N. Am. Chapter of the ACL and Human Language Technology*, pages 173–180. Association for Computational Linguistics.

Ule, T. (2004). Markup manual for the Tübingen partially parsed corpus of written German (TüPP-D/Z). In *Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen*, volume 28, page 2006.

Vidulin, V., Luštrek, M., and Gams, M. (2007). Using genres to improve search engines. In *Proceedings of the International Workshop Towards Genre-Enabled Search Engines*, pages 45–51.

Wan, X. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, pages 235–243. Association for Computational Linguistics.

Wang, S. and Yao, X. (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(4):1119–1130.

Webber, B. (2009). Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682.

Wolters, M. and Kirsten, M. (1999). Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 142–149. Association for Computational Linguistics.

Wu, Z., Markert, K., and Sharoff, S. (2010). Fine-grained genre classification using structural learning algorithms. In *Proceedings of the 48th Annual Meeting of the*

Association for Computational Linguistics, ACL '10, pages 749–759. Association for Computational Linguistics.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In Fisher, D. H., editor, *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers, San Francisco, US.

Yatsko, V. A., Starikov, M. S., and Butakov, A. V. (2010). Automatic genre recognition and adaptive text summarization. *Automatic Documentation and Mathematical Linguistics*, 44:111–120.

Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon University.